
15 Community Indices, Parameters, and Comparisons

Thomas J. Kwak and James T. Peterson

■ 15.1 INTRODUCTION

Understanding assemblages of fishes and how their numbers and compositions change over time and space has long been a fundamental interest of aquatic ecologists and has increasingly become recognized as an important component of fisheries science and management. Whereas much of traditional fisheries management may have focused on single-species approaches, targeting sport or commercial fishes, direct or indirect biotic interactions among fishes may strongly influence target populations. Furthermore, fisheries scientists may frequently be charged with sampling fish populations to detect changes in the aquatic environment, especially those effects related to human activities (e.g., pollution, altered hydrology, or nonnative introductions), and quantitative descriptors of the entire fish assemblage are required for this purpose.

For fish assemblage descriptors to be ecologically relevant, they must be compared over time or among assemblages, and ecologists and biomathematicians have developed procedures to that end. Many of the indices and procedures that we include in this chapter have been developed for use with other taxonomic groups (e.g., plants, invertebrates, and terrestrial animals), or even engineering applications (e.g., communications, Shannon and Weaver 1949), but are equally applicable to the study of fishes. Many have been developed more thoroughly in flowing-water habitats, but the concepts and techniques transfer well to other aquatic systems. In this chapter, we outline, review, and demonstrate quantitative measures and techniques to describe and compare fish assemblages to assist the fisheries scientist in addressing practical research and management objectives.

15.1.1 Definitions

Organisms that occur in a particular place may be classified as a community or an assemblage, and the meaning of these terms varies among ecologists (Morin 1999). The difference between the definitions of these terms lies primarily in the amount and predictability of the interaction among the coexisting organisms. The term

community implies substantial and predictable interaction and may include multiple taxonomic groups such as microflora, plants, and animals, whereas an assemblage is simply the group of species found together and imparts no ecological assumption. Another term, taxocene, is a taxonomically related set of species within a community, such as plants, mammals, or birds (Hutchinson 1978). Thus, a fish community or fish assemblage is a taxocene, yet that term is seldom used. Likewise, a guild is a subset of species that share common resources by similar modes (Root 1967), and although a guild is independent of taxonomy, it is most often a subset of a taxocene. For practical purposes in fish sampling, a fish assemblage is the sum total of the individuals collected at a single sampling location by any single technique or combination of them. For the purposes of this chapter, we employ the commonly applied term fish assemblage to describe the co-occurring fishes in a sample, but we recognize that no fish exists in isolation.

15.1.2 Advantages and Limitations to a Community Approach

The importance of studying fish assemblages, over single species, was evident to early aquatic ecologists (Forbes 1887; Shelford 1929), and today the advantages of a broader ecological approach are obvious and accepted by scientists. However, pragmatic and logistic constraints faced by fisheries scientists do not always allow a holistic perspective. Thus, each investigator must balance the benefit gained in knowledge by a community approach against the additional complexity and effort for each specific application.

The aquatic community is the optimal unit of study as it regulates the flow and storage of energy and materials in the ecosystem. If the fish component is of interest, then the entire fish assemblage is the best unit of study to elucidate the function of this group in the ecosystem. The composition of a fish assemblage is a result of an integration of zoogeography and ecology. Individual fish species vary widely in their morphology, physiology, and tolerance and response to their surroundings. A number of physical factors can limit the ecological success of fish populations, including water quantity, water quality, and physical habitat structure, which in turn set the framework in which biotic interactions occur, such as growth, reproduction, trophic dynamics, and competition (Karr et al. 1986; Fausch et al. 1988; Rabeni and Jacobson 1999). These physical factors may also be quantified by a suite of more proximate measures (e.g., nutrient concentrations, depth profiles, and physical cover) and are further influenced by more broad-scale processes over watersheds and riparian zones. Thus, if any one fish population or guild is limited by a single factor, the effects of other (nonlimiting) environmental influences may not be apparent by merely sampling that fish or subset of fishes.

Fish species of special interest may be atypical in their population dynamics and response to the environment. Some sport and commercial fishes are ubiquitous and tolerant to environmental disturbance (e.g., brown trout, channel catfish, and largemouth bass), and their relative abundance and population dynamics may

depend upon harvest. In contrast, many threatened or endangered fishes are endemic specialists that are extremely sensitive to environmental perturbation (e.g., desert fishes). A widespread, tolerant fish may show no response to habitat degradation or biotic disturbance, whereas a sensitive species may have been extirpated at the earliest signs of perturbation. Thus, single fish species of economic or political importance that are frequently emphasized in fishery surveys or biological assessments may not accurately represent environmental conditions and ecosystem health.

The utility of a community approach is clearly demonstrated in a study by Berkman and Rabeni (1987) to quantify the effects of siltation on stream fishes. They classified fish species from assemblages among sites into guilds based on habitat use, reproductive modes, and feeding behavior. Their guild analysis indicated that species with similar ecological requirements showed a common response to habitat degraded by siltation. Analyses of any single species in their research would likely have been inconclusive, yet the results at the assemblage level yielded strong scientific inference over previous qualitative and anecdotal findings. Further, results examined from a guild approach may allow stronger inference with regard to testing hypotheses about population regulation, as a similar pattern observed among populations within a guild is more conclusive evidence than are trends within a single species.

Valid reasons to forego a community approach in fisheries science also exist. Sampling, sorting, and quantifying all species of a diverse assemblage can be difficult and time consuming, and subsequent data analyses and reporting can be complex. Fish diversity is low in some aquatic ecosystems (e.g., coldwater streams and arctic lakes) and may be dominated by a single species. In such cases, population studies of ecologically important species are reasonable, regardless of practical constraints. Finally, the process of fisheries management and the funding environment for research are strongly governed by economic, sociocultural, and political forces (Krueger and Decker 1999), which may mandate tactical, single-species approaches despite their weaker scientific validity.

Other factors should be considered when contemplating community versus single-species approaches, as no clear criteria exist to guide such decisions. Most fish sampling data are inherently variable over space and time. High variance strongly limits one's ability to detect phenomena statistically, such as the impact of management actions, and can only be overcome by increasing sample size for a given sample design (Chapter 3). Fish assemblage attributes, such as species richness, are generally less variable than are density or abundance estimates for individual species (Peterson and Rabeni 1995). Hence, studies that utilize assemblage data usually require smaller sample sizes to obtain precise estimates and can ultimately be more cost effective than are single-species approaches. The reduction in sample size, however, must be balanced by the additional effort required to process each sample. Thus, any fish sampling protocol should be guided by objectives and scale, specific to the situation, but must be balanced by logistic considerations.

15.1.3 Strategies for Analysis of Community Data

The approach and techniques to employ in analysis of community level data depend upon the objectives of the study and the form and quantity of data. Objectives might be to describe or to compare assemblages. Among descriptive objectives, there may be an emphasis on assemblage structure or ecosystem integrity; comparative objectives may require grouping or ranking of assemblages. Assemblage data may be collected as catch per unit effort or absolute abundance and may be binary (presence–absence), ordinal (ranks), or quantitative (counts), with variable numbers of assemblages and replicates. In Figure 15.1, we present a flow diagram that depicts selection criteria for analytical techniques for community data and may serve as a preliminary guide to the fisheries scientist. Details of criteria, advantages, and shortcomings of each technique are detailed in the appropriate section referenced in the flow diagram. We present example SAS programs (SAS Institute 2004) for performing many of the procedures that we outline in this chapter, analyzing a large-river fish assemblage data set (Box 15.1) along with corresponding results as program output (Boxes 15.4–15.13), but other statistical software applications also perform these procedures. A nonexhaustive list of such software applications is presented in Table 15.1, describing which procedures may be performed using each application. Further, many software applications (e.g., R, SAS, or SPSS) allow more advanced statistical treatment of data depending on the computer programming skills of the user.

15.1.4 Topics Covered

The emphasis of this chapter is on structure, not function; that is, we cover quantitative descriptions of fish assemblage composition and comparison of assemblages rather than describing and understanding community processes and interactions. Assemblage structure is the numerical abundance of each species in the community, and descriptors may include totals or various subtotals of those abundances as well as estimates of biomass. The first step in describing a fish assemblage is designing a sampling protocol to meet specific data requirements, and we provide considerations and suggestions to facilitate that initial process. Once data have been gathered, the process of reducing the resulting data matrix into more meaningful and comparable indices is usually warranted, and we outline those methods in this chapter. We then describe statistical procedures to compare composition among assemblages and conclude with pragmatic suggestions on approaches and interpretation for fisheries scientists. For approaches, methods, and examples that quantify fish community processes, interactions, and forces which, in turn, structure fish assemblages, a topic not covered in this chapter, we refer the reader to Crowder (1990), Gerking (1994), and Matthews (1998); see Krebs (1998), Morin (1999), and Southwood and Henderson (2000) for outstanding general texts on methods for community ecology.

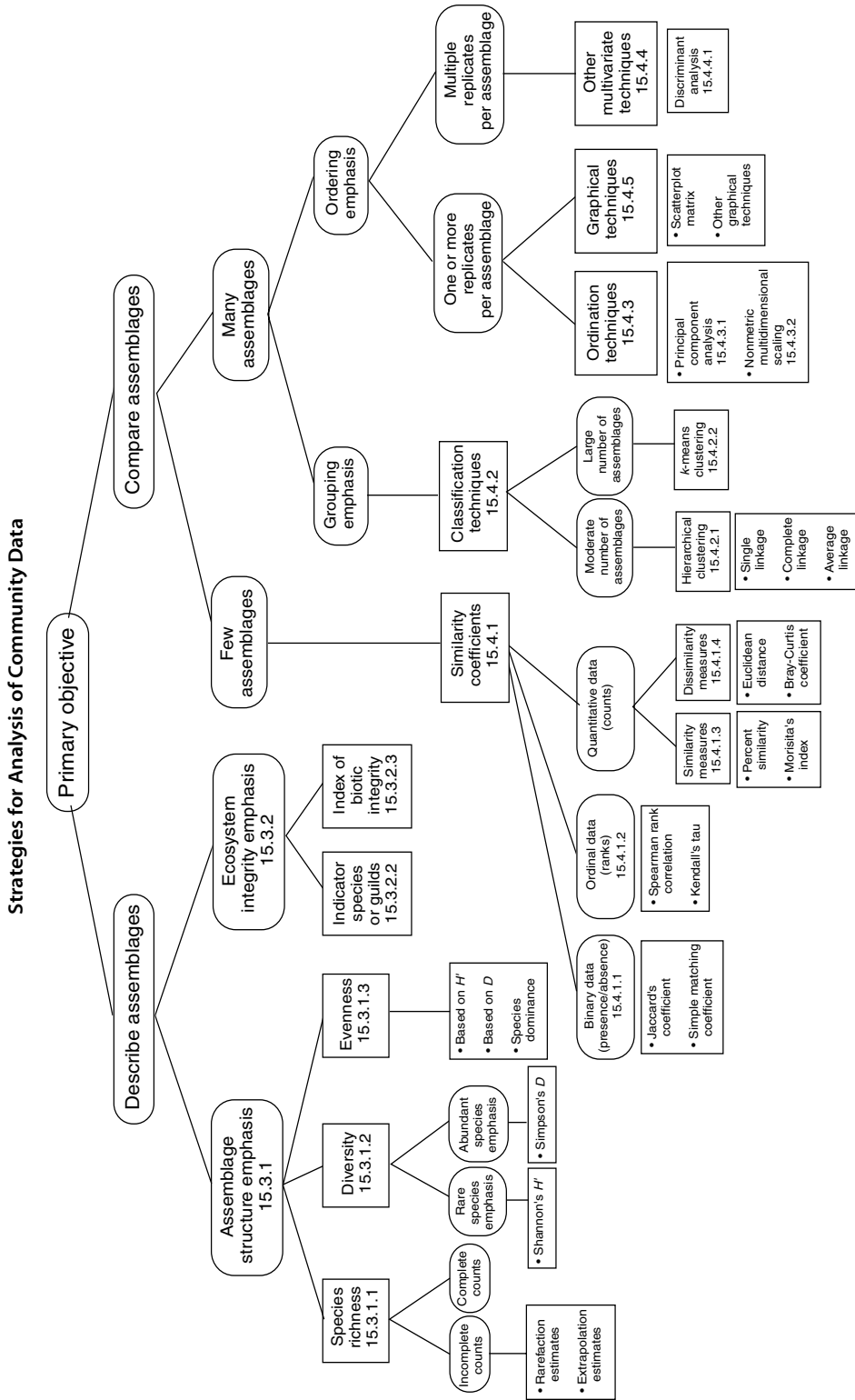


Figure 15.1 Flow diagram depicting selection criteria for techniques available to describe and compare fish assemblages based on objectives and data availability. Objects with rounded borders depict decisions related to objectives or data; those with square borders depict analytical techniques. Numbers refer to appropriate sections within this chapter.

Box 15.1 Sample Data Set and Structural Indices

During a 2-week period in 1988, a survey of the fishes of the Kankakee River, Illinois, was conducted using a boat-mounted electrofisher. Six sites (stations) were each sampled eight times with effort standardized among samples (Peterson 1989; Kwak 1993).

Table The sum of number of individuals from eight samples from each of six sites according to site and species; rare species (occurring in less than 5% of samples) are omitted.

Species and total	Station number					
	1	2	3	4	5	6
Longnose gar	6	7	0	4	26	5
Gizzard shad	164	90	6	6	432	194
Bluntnose minnow	42	33	29	3	44	35
Bullhead minnow	0	0	0	1	15	0
Common carp	13	58	10	14	36	13
Hornyhead chub	0	0	0	0	7	0
Mimic shiner	10	11	10	0	2	0
Redfin shiner	0	22	4	0	0	2
Rosyface shiner	8	89	5	15	8	35
Sand shiner	1	22	4	1	5	1
Spotfin shiner	19	24	3	2	23	8
Striped shiner	45	32	69	14	51	14
Suckermouth minnow	0	0	0	0	4	1
Black redhorse	0	0	4	0	1	0
Golden redhorse	34	0	35	36	9	55
Northern hog sucker	5	2	10	7	2	7
Shorthead redhorse	35	0	8	2	35	22
Quillback	5	2	1	4	17	14
River redhorse	2	0	2	0	0	5
Silver redhorse	8	1	1	3	14	4
Smallmouth buffalo	0	3	0	0	3	0
Brook silverside	4	10	13	9	10	6
Bluegill	2	0	0	2	1	0
Green sunfish	1	42	6	3	7	1
Largemouth bass	0	3	2	0	8	0
Longear sunfish	35	94	26	39	48	37
Orangespotted sunfish	1	0	0	8	31	3
Rock bass	30	3	15	31	27	62
Smallmouth bass	143	59	195	151	165	204
Banded darter	4	0	0	1	0	1
Blackside darter	1	9	1	0	4	0
Johnny darter	0	6	3	0	2	0
Logperch	25	0	51	42	24	7
Slenderhead darter	5	0	6	7	2	2
Total	648	622	519	405	1,063	738

(Box continues)

Box 15.1 (continued)**Table** Indices of fish assemblage structure for six stations. Shannon's index (H') is given by equation (15.3); Simpson's D is given by equation (15.4); and species dominance is given by equation (15.7).

Assemblage structural index	Station number					
	1	2	3	4	5	6
Richness						
Species	26	22	26	24	31	25
Family	7	7	6	7	7	7
Diversity						
Shannon's H'	2.43	2.56	2.29	2.28	2.31	2.24
Simpson's $(1 - D)$	0.864	0.903	0.818	0.821	0.798	0.832
Evenness						
Based on H'	0.746	0.828	0.704	0.718	0.672	0.695
Based on $(1 - D)$	0.898	0.946	0.850	0.857	0.825	0.867
Species dominance (3 species)	0.543	0.439	0.607	0.573	0.610	0.614

Table 15.1 Nonexhaustive list of statistical software applications for analyzing community level data, according to technique. Symbols indicate that the application incorporates some (S), most (M), or all (A) of the corresponding techniques outlined in this chapter.

Statistical software application	Similarity measures	Hierarchical cluster analysis	K-means cluster analysis	Principal component analysis	Nonmetric		
					multi-dimensional scaling	Discriminant analysis	Graphical analysis
BMDP	S	A	A	A	A	A	S
JMP	S	A	A	A		A	M
Minitab	S	A	A	A		A	S
R	S	A	A	A	A	A	M
SAS/STAT	M	A	A	A	A	A	M
S-plus	M	A	A	A	A	A	M
SPSS	S	A	A	A	A	A	S
STATA	S	A	A	A			
Statistica	S	A	A	A	A	A	S
Systat	M	A	A	A	A	A	M

15.2 SAMPLING CONSIDERATIONS AND ASSUMPTIONS

As with all fishery assessments, the analysis and interpretation of fish community indices are significantly influenced by the quality and quantity of data. Fish sampling bias can obscure relationships or, worse, suggest false relations (Bayley and Dowling 1993), and sample variance can affect the ability to detect relationships statistically (Peterson and Rabeni 1995). Standardized sampling protocols help

maintain data quality by ensuring that data are collected in a consistent manner over space and time. The influence of sampling bias and variance on single-species approaches and the importance of standardized sampling have previously been covered elsewhere (e.g., Brown and Austen 1996; Chapters 2 and 3), and most of the principles are applicable to community approaches. There are, however, several considerations unique to sampling entire fish assemblages that we consider here.

When using community indices and comparing communities, the primary assumption is that fish samples are representative of the “true” fish assemblage. That is, the number and types of species caught and their relative abundances accurately reflect those of the fish assemblage occupying the study area (e.g., lake or stream). Fish assemblages, however, are composed of species of different sizes, forms, and behaviors that can affect their vulnerability to capture by any sampling gear. Consequently, samples are influenced by these characteristic differences to varying degrees, resulting in an inaccurate representation of the fish assemblage. Similarly, fish species often use habitats that differ in size, structure, and distribution; hence, the types and allocation (relative amounts) of habitats sampled can also significantly affect the adequacy of the fish assemblage sample. It is important that fisheries scientists identify potential influences on data quality and develop sampling designs that minimize these, so that analyses of fish assemblages will be based on reliable data. Below, we identify some noteworthy influences on the quality of fish assemblage data and discuss methods to minimize their influence and to evaluate the adequacy of sampling designs.

15.2.1 Sources of Sampling Bias

Aquatic ecosystems are defined by characteristic physical, chemical, and biological attributes that may simultaneously influence sampling efficiencies and regulate fish assemblage structure. For example, water depth can affect the efficiency of many fish sampling methods and also can influence the structure of fish assemblages. Consequently, observed differences in the structure of fish assemblages collected in study areas with very different depths could be due to sampling efficiency, assemblage structure, or both. Failing to account for differences in sampling efficiency when comparing locations with different physical characteristics and species assemblages, or among samples collected with different methods, can introduce a systematic error or bias into the data, which can invalidate experiments or observational studies (Hurlbert 1984). To minimize the influence of sampling bias on fish assemblage studies, scientists should collect fishes with the most efficient method or combination of methods for which bias is known and sample under circumstances where catchability is reliable. Thus, fisheries scientists must consider those major factors affecting sampling efficiency and choose the most appropriate gear or combination of gears for their particular sampling situation. When sampling conditions are particularly challenging (e.g., large rivers and reservoirs), fisheries scientists also should consider an analysis based on more qualitative measures (e.g., species presence or rank abundance) or estimate

species abundances using mark–recapture or other methods that can account for sampling efficiency differences.

15.2.1.1 *Species and Body Size*

The efficiency of most sampling gears is influenced by the species and size of fish encountered (Hayes et al. 1996 and references therein). Body shape or morphology can influence a fish's vulnerability to capture. For example, species with cryptic coloration and reduced or absent swim bladders are often difficult to locate when stunned during electrofishing. Species-specific behaviors, such as vertical position in the water column, also affect sampling efficiency. Benthic species (e.g., darters, sculpins, and North American catfishes), particularly those using deepwater habitats, and wide-ranging pelagic species (e.g., temperate basses and herrings) are difficult to sample effectively. Body size, within and among species, is also an important factor affecting sampling efficiency. For most sampling gears, the lowest efficiencies tend to be for the extreme sizes of fish (i.e., very small and large individuals). These sampling biases often result in fish assemblage samples that overrepresent species and sizes that are most vulnerable to sampling. To minimize the influence of these biases, fisheries scientists can develop sampling efficiency models to adjust sampling data for differences in catchability. These estimates, however, require extensive gear evaluations to develop efficiency models. Excluding small fishes and species that are difficult to catch from analyses and using species presence–absence or rank abundances for the assemblage analysis (see sections 15.4.1.1 and 15.4.1.2) also can minimize this source of bias.

15.2.1.2 *Habitat Characteristics*

The physical characteristics of a sampling location can affect the efficiency of most fish sampling gears. The dimensions of a sampling location (e.g., water depth or stream width) can change the capture efficiency of a variety of gears. Sampled areas wider and deeper than the effective catch area (e.g., electrical field size or seine dimensions) can reduce capture efficiency (Bayley and Dowling 1990). In rivers, high current velocities can displace stunned fish from the electrical field before they are captured, facilitate fish escape from seines, and prohibit the use of some passive sampling gears (e.g., gill nets). Similarly, water transparency (color and turbidity) can greatly influence the application and bias of underwater observation techniques and electrofishing and passive-sampling gears. Structures within the sampling area (e.g., vegetation, woody debris, and boulders) can provide a refuge for fishes and can limit sampling efficiencies (Bayley and Dowling 1990; Rodgers et al. 1992). These biases often result in samples that overrepresent those species occupying habitats that are easier to sample and underrepresent those in habitats that impair sampling. Similar to species and size biases, the influence of habitat biases can be minimized by adjusting catch data for differences in sampling efficiency. When sampling efficiency estimates are unavailable, biologists can minimize habitat biases by grouping habitat types into strata (Chapter 3). With such designs, comparisons of fish assemblages should be restricted to similar

habitat types within each stratum. Fisheries scientists should also consider expending greater effort in habitats that are difficult to sample to ensure adequate representation of the fish assemblage in these areas.

15.2.1.3 *Gear Type*

In some instances, using the proper sampling gear can reduce (not eliminate) the influence of species, size, and physical habitat biases on the analysis of fish communities. Thus, selecting the proper sampling gear is one of the most critical components of a fish community study. Electrical gears are among the most efficient and widely used techniques for sampling fish assemblages in relatively shallow waters, such as small- to medium-sized streams and lake and river shorelines. Sampling in deeper waters would likely require the use of active (e.g., dredges or trawls) or passive (e.g., hoop, trap, and fyke nets) techniques designed to sample these habitats more effectively (Hubert 1996). When sampling conditions (habitats) vary considerably within a study area (e.g., large lakes, rivers, and reservoirs), no single sampling gear can adequately sample the entire fish assemblage. Hence, a multi-gear approach, in which the most effective gear(s) is used in each habitat type, can provide the most complete estimate of fish assemblage structure. However, because such estimates can be biased to varying, unknown degrees, they should not be accepted as representative of the “true” fish assemblage unless the effectiveness of each gear can be evaluated.

15.2.2 **Sampling Season**

Season can have a profound influence on fish assemblage structure in many freshwater ecosystems. Fishes often migrate seasonally to fulfill one or more life history requirements (e.g., spawning or juvenile rearing) and to seek refuge during severe environmental conditions (Hall 1972; Schlosser 1982; Bayley and Osborne 1993; Peterson and Rabeni 1996; Grossman et al. 1998). Thus, the structure of the fish assemblage in open systems (e.g., streams and rivers) is likely to vary among seasons where fish can freely migrate to or from study areas. In closed systems (e.g., lakes and ponds), seasonal movements can affect the assemblage structure within habitat types, thereby increasing variance (see section 15.2.3). Similarly, fish movements within a season also can alter variability of assemblage samples. Fish movement is generally greatest during the spring and fall in the northern hemisphere, which can increase sample variance. To avoid the influences of seasonal fish movement, fisheries scientists should limit comparisons of fish assemblages to similar seasons and, if possible, to seasons with the least amount of fish movement (e.g., summer for temperate, warmwater stream fishes).

Fish growth and recruitment to sampling gear also influence seasonal measures of assemblage structure. Young-of-the-year (age-0) fishes—usually the most abundant age-class—are often recruited to sampling gears by the end of their first growing season. This increases the probability of collecting less abundant and difficult to sample species (Gray 1987; Wright 1988), resulting in perceived increases in the number of species during such periods (Peterson and Rabeni 2001).

Eliminating age-0 fishes from an analysis can reduce the influence of seasonal gear recruitment but may also reduce estimates of species richness. When research objectives include analyses of age-0 fish, sampling should be conducted later in the growing season when age-0 fish are larger and more vulnerable to sampling.

15.2.3 Fish Assemblage Sampling Designs

Sampling design is an essential component of fish assemblage studies. Comparisons of fish assemblage structure require data that accurately reflect the true species composition and species' relative abundances. One means of ensuring accurate representation of the species assemblage is through effective sampling design (also see Chapter 3). Fish distribution and assemblage structure are influenced by physical habitat features and resource availability (Gorman and Karr 1978; Schlosser 1982). Thus, designs should ensure that all habitat types in a study area are properly represented. High variance among samples, another factor affecting the accuracy of fish assemblage estimates, is influenced by species-specific characteristics (e.g., behavior) and sampling conditions (e.g., gear type and habitat features) and can be overcome only by increasing sample size. Increasing sample size also improves the likelihood of detecting rare species in the assemblage. The diverse nature of freshwater ecosystems (e.g., habitat types and species) dictates that no single sampling design is best for all community level studies. Rather, the best approach will depend upon the objectives of the study, type of system being studied, and characteristics of the fish assemblage. Here, we discuss two basic approaches to sampling fish assemblages that can be modified to fit most community level studies in freshwater systems. We also strongly encourage scientists to evaluate the adequacy of their particular sampling design to ensure data quality.

15.2.3.1 *Quadrat Sampling*

Quadrat sampling entails dividing a study area into sample units (quadrats) and sampling a random selection of them. With this design, sample variance can be minimized by sampling greater numbers of quadrats. The number of samples required to meet study objectives can be determined with traditional statistical techniques (Chapter 3) and by analyzing species accumulation curves (section 15.3.1.1). Greater efficiency (lower variance) can often be gained by stratifying study areas according to habitat type and randomly sampling quadrats within each stratum.

In addition to sample size requirements, the size of individual quadrats should be considered prior to adopting a quadrat sampling approach. Larger sample units generally contain greater numbers of individuals, which can increase the chances of collecting an individual of another species (Connor and McCoy 1979; Angermeier and Schlosser 1989). To avoid this species-area effect, it is often preferable to maintain a consistent quadrat size among study areas or through time (when monitoring). When study areas differ substantially in size and structure, a single quadrat size could incorporate variable habitat heterogeneity among areas,

which could bias comparisons. For example, pools and riffles generally occur every five to seven stream widths in gravel-dominated streams (Leopold et al. 1964; Gordon et al. 1992). Thus, a single quadrat size based on stream length or area would incorporate a greater number of pools and riffles in smaller streams. In these instances, natural discrete morphological features (i.e., channel units; Hankin and Reeves 1988; Peterson and Rabeni 2001) could be used as sampling quadrats. These natural quadrats, however, should be sampled in proportion to their relative abundance in the study area to ensure proper representation of the fish assemblage.

15.2.3.2 *Constant Ratio Sampling*

Constant ratio sampling involves collecting fishes from a single sample unit, the dimensions of which are scaled relative to the size of the study area. This approach is generally used for stream studies where the size of the sample unit (stream reach length) is proportional to stream width (e.g., station length equals 35 stream widths). The size of the sample unit needed to obtain a representative sample is determined by examining the cumulative catch of species (see section 15.3.1.1) with increasing sample unit size. Thus, this design differs from quadrat sampling in that it attempts to standardize the sampling effort at a single location and time in order to obtain a representative sample of the fish assemblage. The required sample unit size, however, can vary widely among systems due to differences such as habitat characteristics, sampling efficiency, and fish abundance and assemblage structure (Lyons 1992; Angermeier and Smogor 1995; Paller 1995). Hence, no single ratio or proportion will likely be adequate for sampling fish assemblages in all freshwater systems. Additionally, data collected following a constant ratio design cannot be used to make statistical inferences about fish assemblage patterns within a study area because of the general lack of replication (i.e., only one sample collected).

15.2.4 **Data Standardization**

There are innumerable ways to collect and quantify fish assemblage samples. Numerical abundance for each species may be expressed as total catch, relative catch as a proportion of that of all species, catch per unit effort, catch per area, or catch per linear distance; in addition, adjusted or absolute abundance (density or biomass) may be estimated (see Ricker 1975 for examples). Furthermore, the definition of a fish assemblage for quantitative purposes may vary widely (Rahel et al. 1984; Grossman et al. 1990; Matthews 1998).

Investigators may simplify analyses and reduce variable sampling bias by defining the fish assemblage as a subset of the actual sample. This practice may exclude age-0 fish, juvenile fish, or rare species from analyses. Rare species, those found in less than 5% of the collections (Gauch 1982), are generally excluded in most community analyses because (1) it is unlikely that rare species significantly influence the dynamics of the fish community, (2) the occurrence of a rare species may be a random event unrelated to any life history requirement, and (3) many

multivariate statistical techniques are sensitive to rare species, which could distort meaningful, significant trends. However, patterns in occurrence of rare species among assemblages may provide insight into assemblage organization and form the basis of conservation strategies.

Data may also be mathematically transformed to reduce the importance of extreme values (e.g., logarithmic or square root). If data are omitted or transformed for subsequent analyses, it must be justified on some ecological, statistical, or theoretical basis, and data must be modified objectively, based on a systematic criterion. Data manipulation, such as omitting rare species or transformation, will change results of virtually all of the indices and procedures presented in this chapter and should be considered carefully.

Sampling techniques, effort, area, and resulting numerical expressions of fish abundance should be standardized within a study if possible and described in sufficient detail to allow comparisons among studies. Standardizing fish collection techniques and sites may reduce the effects of variable sampling bias associated with gear type, habitat, or time. Standardized data and manipulation will limit erroneous conclusions that may be drawn from statistical artifacts or comparison of incongruous data. However, of utmost importance in community level studies is that a precise definition of the assemblage be provided and data collection and analyses be described in sufficient detail to allow interpretation and additional analyses by others. This detail should include (1) the organisms considered to compose the assemblage (e.g., size or age criteria, rare species criteria, fin fishes, shellfish, other invertebrates, or aquatic herpetofauna); (2) sampling techniques, effort, area and boundaries, and timing; (3) data form and units; and (4) any data manipulation prior to analyses.

■ 15.3 COMMUNITY INDICES—THEIR CHARACTERISTICS AND ESTIMATION

It is indeed appealing and useful to attempt to summarize the abundance data for multiple species in an assemblage into a single number describing assemblage structure. However, this section must begin with a word of caution. Abundant critiques and revisions in the ecological literature suggest that there is no “silver bullet” or perfect community index to serve all purposes. Essentially all of the indices presented below have received criticism for their shortcomings and misapplications (Hurlbert 1971; Washington 1984), and we urge the fisheries scientist to view these indices as relative values with variable precision, accuracy, and reliability. They are most appropriately used to compare assemblage data collected in a standardized manner within a study, rather than as broad, comparative tools among studies or over expanded scales. Index selection should consider statistical robustness, data availability, and specific study objectives. An approach utilizing several indices may prove useful to verify findings and to balance shortcomings of any single index.

Two primary approaches to quantify community structure have been developed by ecologists and applied to fish assemblages. They are the use of (1) community structural indices based directly on field samples and (2) biotic indices

based on the relative abundance of indicator organisms. Both approaches are applicable to describing fish assemblage characteristics and may be related to environmental quality, but biotic indices are especially suited to quantifying ecosystem health or ecological integrity, which may be reduced by pollution, stressful environmental conditions, or habitat degradation and destruction. As such, biotic indices do not directly represent assemblage structure. Structural indices are broad, quantitative descriptors of assemblage structure, and biotic indices are specific parameters based on a subset of indicator organisms within the assemblage, and their applications are not interchangeable.

15.3.1 Structural Indices

The relative abundance of species, other taxa, or other meaningful categorical attribute within an assemblage may be combined into a single measure that is intended to describe the state of the community. The most common of these measures is species diversity, which incorporates the number of species in an assemblage (species richness), as well as the relative abundance of those species (evenness) (Figure 15.2). Whereas such structural indices are usually calculated at the species level, it is equally appropriate to estimate them at any taxonomic or

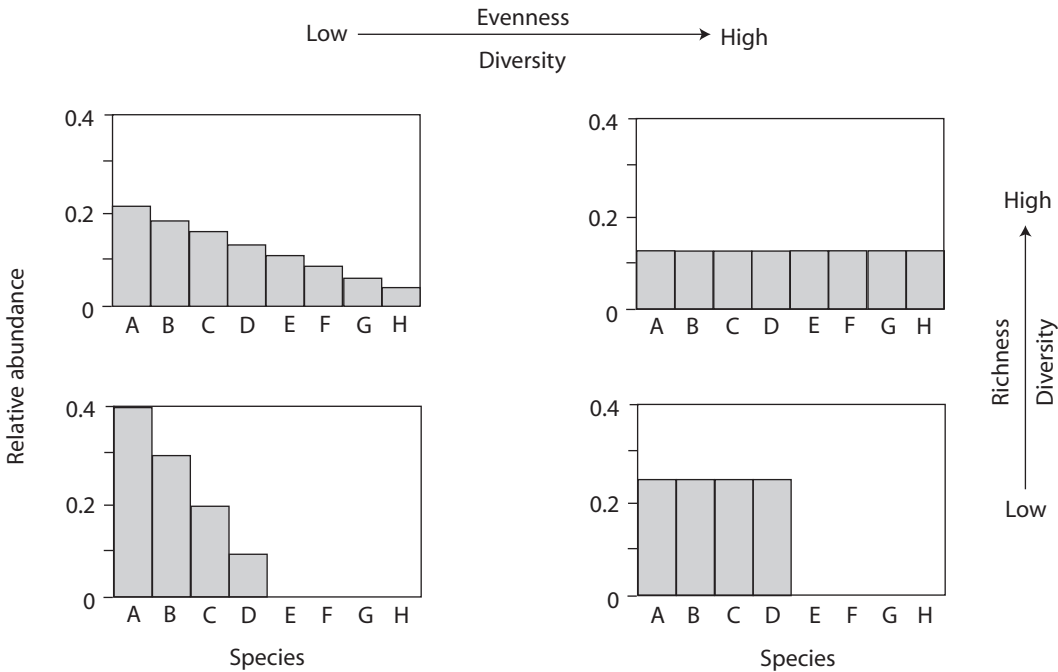


Figure 15.2 The concept of species diversity illustrated by variable relative abundance of species among assemblages. Diversity increases with increases in the number of species (richness) and equitability of distributions among species (evenness).

other hierarchical classification level (Osborne et al. 1980). The level at which to estimate structural indices is often determined by the practical ability to classify organisms. For aquatic invertebrate assemblages, which may comprise species-rich assemblages or contain unresolved taxonomic groups, these indices are often estimated at levels higher than species (e.g., genus or family). But for fish assemblages, the practical level of identification and structural index estimation is usually the species. Most assemblage indices are based on taxonomic classification, as are the examples in this chapter, but it is worth considering other biological and ecological attributes of fishes, in addition to, or in lieu of, taxonomy to describe assemblage structure. Among those that may reveal insightful, functional patterns are life history or morphological traits, habitat affinity at various spatial scales, and tolerance to environmental conditions (Bain et al. 1988; Winemiller and Rose 1992; Poff and Allan 1995; Angermeier and Winston 1999; Quinn and Kwak 2003).

15.3.1.1 *Species Richness*

The simplest and oldest assemblage structural index is species richness—simply a count of the number of species represented in an assemblage. Again, fish assemblage richness may also be expressed at genus, family, or other classification level. For demonstration purposes, an example data set from the Kankakee River, Illinois, is provided in Box 15.1. Among the six sites sampled, fish species richness varied from 22 to 31 species, and family richness varied from six to seven families. The variation between taxonomic levels in this example (nine species versus two families) illustrates the differing utility among levels of classification; structural indices based on very broad class levels are unlikely to provide the resolution necessary to be ecologically relevant.

While the concept of species richness appears simple, that is rarely the case. If an investigator is able to collect or count all individuals of an assemblage in a sampling area, expressing species richness is simple—the count of species present. However, when sampling fishes in aquatic environments, this is rarely the case, and scientists usually collect a sample of the assemblage rather than a complete count. Such samples are incomplete and variably biased by sampling technique and associated influences of sampling habitat, effort, area, and time, as discussed above—all of which affect the ability to sample or detect a species. This limitation is especially important when detecting rare species is a priority. In general, the larger the sample or the greater the number of samples collected, the greater the number of expected species. Consequently, it may be misleading to compare species richness among samples or sites that are based on incomplete counts with varying sample sizes, area sampled, or effort expended. But how can you estimate the number of species not detected? Several approaches to this sampling problem have been developed that are applicable to fish species richness.

Estimating species richness by rarefaction. Rarefaction is a statistical method to compare species richness among assemblage samples of different sizes (i.e., different numbers of individuals per sample). This procedure was first developed by Sanders (1968) to compare marine benthic assemblages and was later corrected for an

error by Hurlbert (1971) and Simberloff (1972). It mathematically “rarefies” a large sample of known species richness to estimate what richness would be for samples of fewer individuals. If rarefaction is performed for a number of sample sizes, a rarefaction curve can be constructed for that assemblage to serve as a tool for comparing species richness among assemblages for equal sample sizes. This process thus reveals differences in species richness among assemblages, independent of sample size.

The rarefaction algorithm assumes a hypergeometric distribution of the species–abundance relationship as

$$E(S_n) = \sum_{i=1}^S \left[1 - \frac{\binom{N-N_i}{n}}{\binom{N}{n}} \right], \quad (15.1)$$

- $E(S_n)$ = expected species richness of a random subsample;
 S = total number of species in the collection;
 N = total number of individuals in the collection;
 N_i = number of individuals of species i ;
 n = number of individuals in the random subsample; and
 $\binom{N}{n}$ = number of combinations of n individuals that can be selected from a sample of N individuals, or $N!/n!(N-n)!$.

An example calculation of expected species richness of a smaller sample is presented in Box 15.2, and the algorithm to estimate the variance of $E(S_n)$ may be found in Heck et al. (1975). Once a large sample has been rarefied, expected species richness from smaller samples of that assemblage can be compared to samples of equal abundance from other assemblages to compare richness among assemblages, independent of sample size. Plotting rarefaction curves (see Box 15.2) of multiple assemblages on a single plot is a useful means to compare species richness. Furthermore, if fish density (number per area) of an assemblage is estimated (Chapter 8), expected species richness as a function of sampling area can also be plotted (i.e., species density curve; see Gotelli and Graves [1996] for an example).

Rarefaction can be a useful fisheries or ecological tool (see examples by Glowacki and Penczak [2000] and Quinn and Kwak [2003]). Management strategies, ecological assessment, and hypothesis testing require information on species richness and diversity, independent of sampling size, area, or effort. Fish assemblages may be compared using rarefaction over time or among locations, and associated precision may be estimated as confidence intervals. Rarefaction may also be employed in the development of monitoring programs to determine the sufficient sample size and effort to detect an acceptable proportion of species present.

Box 15.2 Estimation of Species Richness by Rarefaction

A cumulative sample of fishes from station 1 of the Kankakee River, Illinois, included 648 individuals representing 26 species (see Box 15.1). Below, we estimate the expected species richness from a sample of 100 individuals.

From equation (15.1),

$$E(S_n) = \sum_{i=1}^S \left[1 - \frac{\binom{N-N_i}{n}}{\binom{N}{n}} \right]$$

$$E(S_{100}) = \left[1 - \frac{\binom{648-6}{100}}{\binom{648}{100}} \right] + \left[1 - \frac{\binom{648-164}{100}}{\binom{648}{100}} \right] + \left[1 - \frac{\binom{648-N_i}{100}}{\binom{648}{100}} \right] + \left[1 - \frac{\binom{648-5}{100}}{\binom{648}{100}} \right]$$

(longnose gar) (gizzard shad) (23 other species) (slenderhead darter)

The summation term for longnose gar is calculated as

$$\frac{\binom{648-6}{100}}{\binom{648}{100}} = \frac{642!}{100!(642-100)!} = 1.7668 \times 10^{119}$$

$$\frac{\binom{648}{100}}{\binom{648}{100}} = \frac{648!}{100!(648-100)!} = 4.8506 \times 10^{119}$$

$$\left[1 - \frac{1.7668 \times 10^{119}}{4.8506 \times 10^{119}} \right] = 0.6358.$$

Thus,

$$\begin{aligned} E(S_{100}) &= \text{longnose gar term} + \text{gizzard shad term} + \text{total of 23 other species terms} \\ &\quad + \text{slenderhead darter term} \\ &= 0.6358 + 1.0 + 15.4213 + 0.5687 \\ &= 17.626 \text{ species.} \end{aligned}$$

Therefore, we would expect a random sample of 100 fish from station 1 to include about 18 species. We may then calculate expected species richness for a number of other smaller sample sizes by repeating the calculation above, varying n to develop a rarefaction curve. For station 1, some values of $E(S_n)$ are $E(S_{50}) = 14.040$; $E(S_{200}) = 21.110$; $E(S_{300}) = 22.940$; $E(S_{400}) = 24.107$; $E(S_{500}) = 24.975$; and $E(S_{600}) = 25.693$, resulting in the rarefaction curve below.

(Box continues)

Box 15.2 (continued)

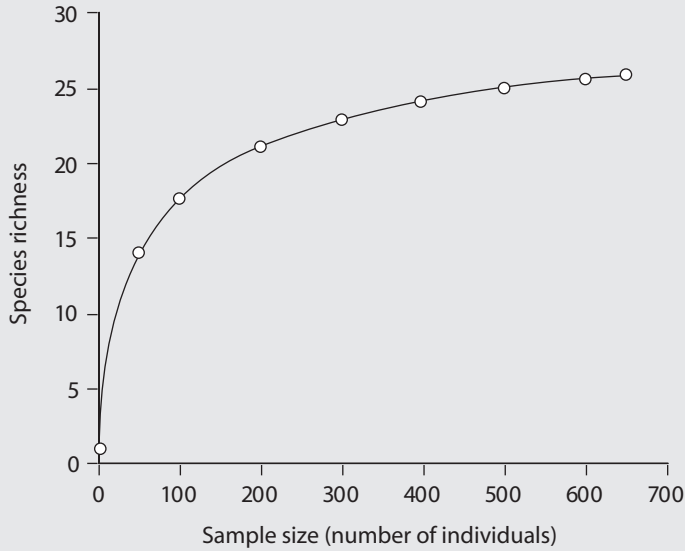


Figure Rarefaction curve to estimate expected species richness of the fish assemblage at station 1 (from Kankakee River, Illinois; Box 15.1) based on sample size. Species richness cannot be estimated for sample sizes that exceed those of the data upon which the curve has been developed.

Note that in the case where $(N - N_i)$ is less than n , then $\binom{N - N_i}{n} = 0$ by definition (no combination of a greater number of individuals can be chosen from a set of fewer individuals). In this case

$$\left[1 - \frac{\binom{N - N_i}{n}}{\binom{N}{n}} \right] = 1 .$$

Unfortunately, the equations above require calculations of very large numeric values (e.g., 648!, as above) that cannot be processed directly on a personal computer. Numbers greater than 170! cannot be stored in floating-point, double-precision arithmetic on a typical personal computer. However, we can perform the calculations on lower values by using the natural logarithm of the gamma function, $\Gamma(x)$ (GAMMALN function in Microsoft Excel), where the factorial of any integer (x) may be calculated as

$$x! = e^{\log_e \Gamma(x+1)} .$$

Equation (15.1), expressed using the gamma logarithm, is

$$E(S_n) = \sum_{i=1}^S [1 - e^{\{\log_e \Gamma(N - N_i + 1) - \{\log_e \Gamma(n + 1)\} + \{\log_e \Gamma(N - N_i - n + 1)\}\} - \{\log_e \Gamma(N + 1)\} - \{\log_e \Gamma(n + 1)\} + \{\log_e \Gamma(N - n + 1)\}}] .$$

(Box continues)

Box 15.2 (continued)

Each summation term is expressed as a formula in Excel as

$$1 - (\text{EXP}((\text{GAMMALN}(N - N_j + 1) - (\text{GAMMALN}(n + 1) + \text{GAMMALN}(N - N_j - n + 1))) - (\text{GAMMALN}(N + 1) - (\text{GAMMALN}(n + 1) + \text{GAMMALN}(N - n + 1)))))$$

The Excel formula for the longnose gar term in the example above would be

$$1 - (\text{EXP}((\text{GAMMALN}(648 - 6 + 1) - (\text{GAMMALN}(100 + 1) + \text{GAMMALN}(648 - 6 - 100 + 1))) - (\text{GAMMALN}(648 + 1) - (\text{GAMMALN}(100 + 1) + \text{GAMMALN}(648 - 100 + 1)))))$$

These tedious computations are best carried out in a spreadsheet application or a specifically developed computer program, such as that provided by Krebs (1998).

Rarefaction has several limitations and assumptions that must be considered in its use and interpretation of results. Rarefaction curves may not be extrapolated beyond the number of individuals in the largest sample. Thus, we only address the question of undetected species for smaller samples. Rarefaction should be applied only to samples from similar habitats using similar sampling techniques. For example, the rarefaction curve developed from sampling fishes of the Kankakee River by means of a boat-mounted electrofisher (Box 15.2) should not be compared with samples from that river collected using other gears or from other water bodies, which we know support different species diversities. Rarefaction assumes a random distribution of individuals, which is rarely true for fishes, and it does not incorporate information about species identity or relative abundance among species.

Estimating species richness by extrapolation. Rarefaction estimates species richness for smaller samples of individuals, but an investigator may wish to estimate the total number of species in an assemblage (i.e., how many species remain undetected?). Careful extrapolation is required to address such questions. A simple technique to extrapolate species richness beyond the boundaries of empirical data is to develop a species accumulation curve. In this procedure, the cumulative number of species collected is plotted against increasing numbers of combined samples of equal effort or area. If samples are quadrat samples within a larger area, their sequential order on the plot should be random; if the samples are a time series from the sample site, they may be applied sequentially or randomly. Various regression techniques have been used to model the resulting relationship, but the linear regression of cumulative species as a function of the logarithm (base 10 log-linear model) generally performs well on empirical data and is simple to apply (Palmer 1990). This plot and resulting regression model may be used to estimate species richness, coinciding with larger numbers of samples (greater effort or sampling area).

We present the species accumulation curve for eight sequential fish assemblage samples from station 1 of the Kankakee River, Illinois, as an example (Figure 15.3, see Box 15.1 for summed data). After sampling the same area eight times with equal effort during a 2-week period, we collected 26 species. However, cumulative species richness increased from 18 species after our first sample to that cumulative total. Examination of the species accumulation curve, extrapolated to 100 samples, suggests that total species richness for that area is greater and that additional sampling would have increased our species count (35.6 species, Figure 15.3b). For example, if resources were available to collect twice as many samples (16), we can use the species accumulation regression function (Figure 15.3) to estimate that we could expect to collect 2–3 more species at that site (28.5 species).

The log-linear model is nonasymptotic; that is, species richness will continue to increase with samples. There are several other asymptotic models available that may be applied to species accumulation curves and parametric and nonparametric estimators to extrapolate estimates of species richness to a maximum number of species; these are reviewed by Colwell and Coddington (1994). For a single sample of an assemblage, they identified an eloquent, nonparametric estimator by Chao (1984) as the best for estimating total species richness (S_{total}) as

$$S_{\text{total}} = S_{\text{obs}} + (a^2/2b), \quad (15.2)$$

where S_{obs} = observed number of species in a sample, a = number of species represented by a single individual in the sample, and b = number of species represented by exactly two individuals in the sample.

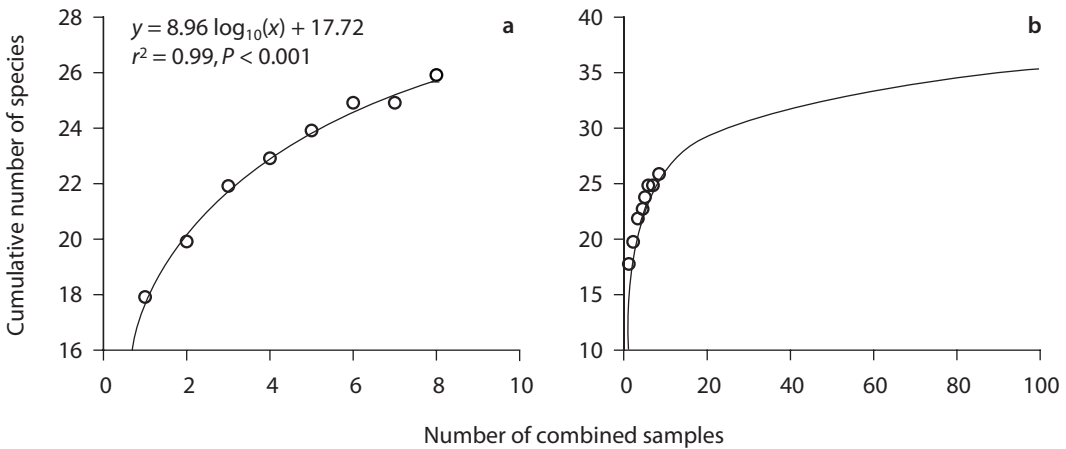


Figure 15.3 Species accumulation curve for eight sequential fish samples collected from station 1 on the Kankakee River, Illinois (Box 15.1), describing the relationship of cumulative species richness (y) and number of samples (x) fitted to a log-linear function. Left panel (a) is the curve within the range of empirical data; right panel (b) is the same relationship extrapolated to 100 samples.

Thus, for our Kankakee River fish assemblage example (station 1; Box 15.1), the total estimated species richness for that site would be 30 species ($S_{\text{obs}} = 26$, $a = 4$, $b = 2$). However, our species accumulation function for that site (Figure 15.3) suggests that about 24 samples (if $y = 30$, $x = 23.47$) would be required to detect 30 species. A variance estimator for S_{total} and application of equation (15.2) to presence–absence data are found in Chao (1984). Recently, ecologists have applied models, originally developed to estimate population size and related parameters, to estimate species richness for communities that include species with varying detection probabilities; such models require multiple samples but may offer advantages over other extrapolation techniques (Boulinier et al. 1998; Bayley and Peterson 2001).

Similar limitations and assumptions, noted for rarefaction above, apply to extrapolation techniques for estimating species richness. However, we urge extreme caution in applying any of these techniques or any statistical procedure that derives estimates by extrapolation beyond the boundaries of empirical data. The nature and form of a relationship may change at wider ranges of variables. This is dangerous territory indeed, and caution and common sense must be exercised to avoid reporting and accepting erroneous and invalid findings. Extrapolation procedures are best employed to develop hypotheses and management scenarios for testing rather than to be used as a basis for critical management decisions.

15.3.1.2 Diversity

Diversity indices combine information on the number of species in an assemblage (richness) and their relative abundance (evenness). Unfortunately, there is no correct means of assigning proportional weighting between these two components, and thus dozens of diversity indices have been developed and applied by ecologists seeking to improve on previous forms (Hurlbert 1971; Washington 1984). What, then, does a diversity index convey? A diversity index is a parameter describing assemblage structure, but any relationship to ecological function, such as productivity or stability, remains unclear. Diversity indices have been criticized for lack of biological relevance and should be considered only one of many tools available to describe assemblage structure—they are not a substitute for in-depth examination (Hurlbert 1971; Pielou 1975; Washington 1984).

We provide equations and examples of two common diversity indices applied at the species level for fishes—Shannon’s H' and Simpson’s D . Shannon’s index of diversity (Shannon and Weaver 1949) has endured ongoing criticism yet remains widely used in biology, and it is the most widely applied diversity index in aquatic systems (Washington 1984). Despite its theoretical and ecological shortcomings, its use is probably justified as a comparative index until such time that a more suitable alternative becomes accepted. It was independently developed by Shannon and Wiener at about the same time and is often referred to as the Shannon–Wiener index or function. Shannon’s index (H') is based on information theory and is defined as

$$H' = -\sum_{i=1}^S (p_i) (\log_e p_i), \quad (15.3)$$

where s = number of species, and p_i = proportion of the total sample represented by the i th species.

Another less common, but among the simplest, diversity (or concentration) index is Simpson's D (Simpson 1949). Simpson's index of diversity is based on the notion that diversity is inversely related to the probability that two individuals sampled at random from an assemblage will be of the same species. Thus,

$$D = \sum_{i=1}^s (p_i^2), \quad (15.4)$$

D = Simpson's measure of concentration,

$1 - D$ = Simpson's diversity index,

s = number of species, and

p_i = proportion of the total sample represented by the i th species.

Shannon's index is sensitive to changes in rare species in the community and is considered a type I diversity index, whereas Simpson's index is influenced to a greater extent by abundant species and is a type II index (Peet 1974; Krebs 1998). Thus, selection of a diversity index should not be an arbitrary process and may vary with specific study objectives and investigator interests. Shannon's, Simpson's, and other diversity indices have been represented by various forms that are based on the original theory proposed (Pielou 1975; Washington 1984; Krebs 1998), so investigators should report the exact algorithm used to compute the index rather than simply citing a reference. Example calculations of species diversity indices for the Kankakee River fish assemblage are presented in Box 15.3.

There are many alternative statistical and practical methods to describe diversity. Shannon's and Simpson's diversity indices are nonparametric measures that imply no assumption about the species abundance distribution of an assemblage. An alternative approach used by community ecologists to describe diversity is to use statistical sampling theory to fit distribution models to species abundance data; examples of these include the logarithmic series, lognormal, geometric series, uniform, and broken-stick distributions (Gotelli and Graves 1996; Krebs 1998). Because of the complexity and lack of theoretical or biological justification for these statistical distribution approaches, nonparametric indices, such as Shannon's and Simpson's have become more widely applied in aquatic science (Washington 1984; Krebs 1998).

15.3.1.3 Evenness

Evenness is a measure of the equitability in relative abundance among species. To report only diversity as an assemblage structural index confounds the effects of species richness and evenness; thus, it is appropriate to report richness, diversity, and evenness when describing fish assemblage structure. There are many approaches to quantifying evenness, but the most common is to express it as a proportion of estimated diversity relative to the corresponding maximum diversity for the specific number of species and sample size.

Box 15.3 Calculation of Species Diversity, Evenness, and Dominance.

Below, we estimate species diversity and evenness of the fish assemblage sample from station 1 of the Kankakee River, Illinois, which included 648 individuals representing 26 species (Box 15.1).

Shannon's Diversity Index (H')

From equation (15.3),

$$H' = -\sum_{i=1}^s (p_i)(\log_e p_i).$$

$$\begin{aligned} H' &= -[(0.009)(\log_e 0.009) + (0.253)(\log_e 0.253) + (p_i)(\log_e p_i) + (0.008)(\log_e 0.008)] \\ &\quad \text{(longnose gar)} \quad \text{(gizzard shad)} \quad \text{(23 other species)} \quad \text{(slenderhead darter)} \\ &= -[(-0.042) + (-0.348) + (-2.002) + (-0.039)] \\ &= 2.431 \text{ nats/individual.} \end{aligned}$$

Units of expression for H' are "nats per individual" (if calculated using \log_e , from information theory; Pielou 1975), but it is usually reported as a unitless index value. Although use of \log_e (as above) has become convention for calculating H' , any logarithm base may be applied (e.g., \log_2 or \log_{10}), as resulting index values are easily converted (see Krebs 1998 for multipliers).

Evenness Based on Shannon's Index (J')

From equation (15.5), with $H' = 2.431$ (above) and 26 species (s),

$$\begin{aligned} J' &= \frac{H'}{H'_{\max}} = \frac{H'}{\log_e s} \\ J' &= \frac{2.431}{\log_e 26} = \frac{2.431}{3.258} = 0.746. \end{aligned}$$

All measures of evenness range from 0 to 1.0 and are unitless proportions.

Simpson's Diversity Index ($1 - D$)

From equation (15.4),

$$D = \sum_{i=1}^s (p_i^2).$$

$$\begin{aligned} D &= 0.009^2 + 0.253^2 + p_i^2 + 0.008^2 \\ &\quad \text{(longnose gar)} \quad \text{(gizzard shad)} \quad \text{(23 other species)} \quad \text{(slenderhead darter)} \\ &= 0.00008 + 0.06401 + 0.07197 + 0.00006 \\ &= 0.13612, \text{ and} \\ 1 - D &= 0.86388. \end{aligned}$$

This result ($1 - D$) is the probability, without units, that two individuals selected randomly from this sample will be different species.

(Box continues)

Box 15.3 (continued)**Evenness Based on Simpson's Index (V')**

From equation (15.6), with $1 - D = 0.8638$ (above) and 26 species,

$$V' = \frac{1 - D}{(1 - D)_{\max}} = \frac{1 - D}{1 - 1/s}, \text{ and}$$

$$V' = \frac{1 - 0.1361}{1 - 1/26} = \frac{0.8639}{0.9615} = 0.8985.$$

Dominance

Based on the three most numerous species, from equation (15.7), dominance (D) is given by

$$D_3 = \sum_{i=1}^3 p_i.$$

$$D_3 = \begin{array}{ccccccc} 0.253 & + & 0.221 & + & 0.069 \\ \text{(gizzard shad)} & & \text{(smallmouth bass)} & & \text{(striped shiner)} \\ = 0.543. \end{array}$$

Maximum value for species dominance would be 1.0. This result is a proportion without units but may also be appropriately expressed (multiplied by 100) as a percentage.

For Shannon's diversity index (H'), the corresponding index of evenness (J') is calculated as

$$J' = \frac{H'}{H'_{\max}} = \frac{H'}{\log_e s}, \quad (15.5)$$

where $H'_{\max} = \log_e s =$ maximum possible value of Shannon's index, and $s =$ number of species.

There is disagreement on a theoretical upper limit for Shannon's index, but in practice, it rarely exceeds 5.0 for biological assemblages (Washington 1984).

The analogous equation to calculate evenness (V') for Simpson's diversity index ($1 - D$) is

$$V' = \frac{1 - D}{(1 - D)_{\max}} = \frac{1 - D}{1 - 1/s}, \quad (15.6)$$

where $(1 - D)_{\max} = 1 - 1/s =$ maximum possible value of Simpson's index, and $s =$ number of species.

The maximum value that Simpson's index may attain is nearly 1.0. Because many variants of Shannon's and Simpson's diversity indices have been proposed, there are an equal number of corresponding algorithms to estimate evenness associated with those measures of diversity (Pielou 1975; Washington 1984; Krebs 1998). For this reason, we suggest consistently reporting the explicit equation used to estimate evenness, as well as that for diversity. Example calculations of species evenness based on Shannon's and Simpson's diversity indices for the Kankakee River fish assemblage are presented in Box 15.3.

Another simple assemblage structural index related to evenness is species dominance, which may be expressed as the relative abundance of a subset of the most numerous species. For example, the proportion of the assemblage composed of the two or three most abundant species would, in general, be inversely related to evenness. The equation to calculate species dominance for the three most abundant species (D_3) is simply

$$D_3 = \sum_{i=1}^3 p_i, \quad (15.7)$$

where p_i = proportion of the total sample represented by the i th species. Species dominance may be estimated for a variable number of dominant species (usually two to three). An example calculation is presented for the Kankakee River fish assemblage in Box 15.3.

15.3.2 *Biotic Integrity Indices*

The concept of using indicator organisms as descriptors of environmental quality may date back centuries, but it was not until the early twentieth century that it became formalized. The "Saprobien system," developed by Kolkwitz and Marsson (1908) in Europe, delineated zones of organic enrichment and classified animal species that occupy them. That early biotic index was later applied to river systems and modified (Chandler 1970), and this led to the prolific development of a variety of biotic indices for aquatic invertebrates that appears to continue without consensus (Washington 1984; Rosenburg and Resh 1993). While indicator species, such as common carp or salmonid species, have been recognized in fisheries science for decades, and other multimetric indices have been proposed (e.g., Gammon's [1976] index of well being), the development and first widespread application of a formal biotic index based on fishes is attributed to James Karr and his colleagues (Karr 1981; Karr et al. 1986).

15.3.2.1 *Rationale*

Biotic integrity of an ecosystem is the ability to support and maintain a balanced, integrated, adaptive community with assemblage characteristics and functional organization similar to a natural habitat in the region that has not been impaired by human activities (Karr et al. 1986). Systems with biotic integrity are more resistant and resilient to natural disturbances and may withstand substantial human influences. Ecological integrity integrates aspects of the chemical and physical

state of the ecosystem with the biological. Whereas aquatic systems with ecological integrity may support productive fisheries or other products and services, ecological integrity is not necessarily correlated with productivity or diversity.

Biotic indices are developed to describe or quantify ecological integrity based on known or suspected relationships between indicator organisms and their environment and may also include assemblage structural indices. Indicator organisms may be selected because they are particularly sensitive or tolerant to environmental degradation, and both types may be incorporated into a single biotic index. Effective biotic indices cannot be universal; as fauna and environmental stresses change regionally, so will suitable indicator organisms. Thus, a biotic index developed for a specific region and environmental stressors may require modification for a different fauna and environmental relationships. Unfortunately, biotic indices are often applied uncritically to systems other than those for which they were developed.

The concept and practice of biotic indices have been widely lauded and criticized on various grounds (Suter 1993; Davis 1995; Simon 1999b). In general, criticisms include a perceived lack of ecological meaning, predictability, diagnostic power, and direct application to water resource regulation. Such criticisms apply to many of the multimetric indices or multivariate techniques covered in this chapter and have been refuted by those successfully applying biotic indices. The differences between opponents and proponents are primarily philosophical and can be overcome by caution and reason in application of techniques and interpretation of results.

15.3.2.2 *Indicator Species and Guilds*

Fishes are especially well suited as taxa to indicate environmental quality (Karr et al. 1986; Simon 1999b). They occur in all but the most degraded waters; they can accurately reflect environmental conditions at multiple scales; life history and geographic distribution information is extensive for many species; and effective techniques are available to collect them. Finally, fishes are relatively more visible, understood, and valued by regulators, politicians, and the general public than are other aquatic organisms.

The indicator fish approach is simple and easily applied without intensive data needs or analysis. Indicator fishes or guilds may be particularly sensitive or tolerant to environmental degradation. The application is more biologically relevant when indicator guilds are used because the effect of their occurrence may imply ecological function, such as feeding or reproduction, rather than specific responses of individual species. Examples of fish guilds to be considered are those based on feeding and trophic relations (Gerking 1994), reproduction (Balon 1975), or habitat (Grossman and Freeman 1987; Bain et al. 1988). Furthermore, higher levels of fish taxa (e.g., families or genera, such as Salmonidae or darters of genera *Ammocrypta*, *Etheostoma*, and *Percina*) may be considered indicator taxa.

Disadvantages of the indicator fish approach lie primarily in its subjectivity and ecological basis. Although several lists partitioning fish guilds exist (e.g., Balon 1975; Karr et al. 1986; Halliwell et al. 1999; Simon 1999c), standard criteria for

guild delineation and selection of appropriate guilds are lacking. Another problem is that mechanisms unrelated to ecological integrity may influence occurrence or ecological success of a fish taxon or guild; these may include zoogeography, biotic interactions, or harvest (Fausch et al. 1990). Further complicating the use of indicator fishes or guilds is that responses to environmental conditions in fishes can vary with space, time, and type or degree of environmental stress, which could confound conclusions among ecosystems or years.

15.3.2.3 *Index of Biotic Integrity*

Since its conception and original development for wadeable, warmwater streams in the Midwestern United States, the index of biotic integrity (IBI, Karr 1981; Karr et al. 1986) has been modified, as intended by the original authors, and applied to virtually all other aquatic ecosystems, including coldwater streams, large rivers, lakes, estuaries, and highly modified habitats, and to various regions of the United States (Simon 1999a). Today, the IBI is widely applied and serves the function of a conceptual and procedural framework for assessing biological integrity based on fish assemblages rather than a prescribed, specific protocol.

The IBI was designed as a composite index to assess biological integrity of aquatic ecosystems by integrating attributes of the fish assemblage, population, and individual by means of relative abundance of species and condition of individuals in a representative sample of the assemblage. Although assemblage structural indices (section 15.3.1) utilize relative abundance data and may reflect ecological conditions in some applications, they were not conceived and designed for that function. The primary advantage of the IBI is that it was specifically developed and refined, based on ecological relationships of fishes, to describe ecological integrity and anthropogenic alterations of aquatic ecosystems.

The original IBI framework included 12 metrics that describe various aspects of fish species composition, trophic composition, abundance, and condition (Table 15.2), but metrics have been omitted, augmented, or modified in applying the IBI to other regions, habitats, and specific ecosystems, usually retaining the original ecological framework (Miller et al. 1988). Increasingly, metrics are developed systematically for a region based on metric variability and empirical relationships (Hughes et al. 1998; Angermeier et al. 2000). A number rating, or score (5, 3, or 1), based on ecological expectations is assigned to each metric, and metric scores are summed to yield a composite index score. The IBI scores may then be compared directly or ranges may be assigned to successive categorical integrity classes from very poor to excellent.

Species composition. The metrics describing species composition were intended to characterize biological integrity through measures of fish species diversity and occurrence of relatively tolerant and intolerant species. Species richness or relative abundance may be modified to reflect that of native fishes in areas affected by nonnative fishes. Occurrence and relative abundance of specific families or taxa may also vary among regions and should include species-rich groups with wide geographic distributions and include one primarily benthic taxon and one nonbenthic taxon (Karr et al. 1986). Species considered intolerant (or tolerant)

Table 15.2 Generalized fish assemblage metrics and scoring criteria for the index of biotic integrity (IBI) applied to streams (modified from Karr et al. 1986). Scores are assigned to each metric based on the sample deviation from that expected from a relatively undisturbed reference system.

Attribute category and metric	Scoring criteria		
	5 (highest integrity)	3	1 (lowest integrity)
Species richness and composition			
Total number of fish species	Expectations vary with stream size, region, and basin (see section 15.3.2.4 for discussion)		
Number and identity of darter species	Expectations vary with stream size, region, and basin (see section 15.3.2.4 for discussion)		
Number and identity of sunfish species	Expectations vary with stream size, region, and basin (see section 15.3.2.4 for discussion)		
Number and identity of sucker species	Expectations vary with stream size, region, and basin (see section 15.3.2.4 for discussion)		
Number and identity of intolerant species	Expectations vary with stream size, region, and basin (see section 15.3.2.4 for discussion)		
Percent individuals as green sunfish	<5%	5–20%	>20%
Trophic composition			
Percent individuals as omnivores	<20%	20–45%	>45%
Percent individuals as insectivorous cyprinids	>45%	20–45%	<20%
Percent individuals as piscivores	>5%	1–5%	<1%
Fish abundance and condition			
Number of individuals sampled	Expectations vary with stream size, region, and basin (see section 15.3.2.4 for discussion)		
Percent individuals as hybrids	0	>0–1%	>1%
Percent individuals diseased or with anomalies	0–2%	>2–5%	>5%
Total IBI score (sum of 12 metrics)	60		12
Integrity class	Excellent – Good – Fair – Poor – Very Poor		

should include only 5–10% of the species that are most (or least) sensitive to human alteration of ecosystems.

Trophic composition. Trophic composition metrics are based on the premise that alterations in food resources and productivity, influenced by water and habitat quality, are reflected in the trophic structure of the fish assemblage. This extends the attributes of the IBI to other trophic levels and organisms. As habitat degrades, food resources fluctuate more, and omnivores may replace more specialized feeders. The presence of piscivores, or other top carnivores, indicates a more complex food web. For classification purposes, omnivores are defined as species that consume significant quantities of both plant and animal material, including detritus (Karr et al. 1986). In regions where insectivorous cyprinids are not common, other insectivorous fish taxa or other specialized feeder may be substituted.

Fish abundance and condition. Metrics describing fish abundance and condition were designed to incorporate population and individual level effects of environmental degradation. Obviously, the number of individuals in the sample will be dependent on effort, and thus, fish abundance must be standardized to units of catch per effort for comparison among sites (see section 15.2.4 and Chapter 7). While a metric based on fish numbers accounts for numbers of trophic links, a

metric based on fish biomass may also be incorporated to account for the magnitude of trophic transfer and energy sequestered. The metrics for hybrid individuals and for disease and anomalies are among those most difficult to apply and are frequently omitted or replaced (Miller et al. 1988). One strength of the IBI is that it is a simple field assessment, but if fish need to be preserved for later analysis, that advantage is diminished. Metrics related to nonnative species abundance or reproductive guilds have been substituted for the hybrid metric to represent a similar ecological rationale.

15.3.2.4 *Spatial Influences and Reference Systems*

The discussion above on IBI attribute categories emphasizes the practice of refining or replacing metrics for application to specific regions, and, likewise, metric expectations should be similarly scaled according to stream size and compared with least-disturbed, reference conditions. This practice will reduce the influence of confounding factors that are unrelated to human influences on fish assemblages and will improve the relationship of index scores to ecological integrity. Such considerations apply as well to comparison and interpretation of structural indices (section 15.3.1) and other methods to compare fish assemblages (section 15.4) and are discussed below.

Regional influence. The variation in fish fauna and assemblages among regions may be influenced by broad-scale factors, such as geological phenomena, river basin boundaries, historical biogeography, glaciation, and evolution (Matthews 1998). Thus, comparative use of the IBI and, in most applications, fish assemblage structural indices or multivariate techniques, should occur within a geographic region. Typical spatial frameworks for delineating regions in this context are by ecoregion, drainage basin, or other related hierarchical division (Bailey 1995; Omernick 1995; Omernick and Bailey 1997; Angermeier et al. 2000).

Stream size and longitudinal influence. Numerous studies on the distribution of stream fishes from headwater reaches downstream to large rivers suggest common patterns of change in fish assemblages along a longitudinal gradient. In most systems, species richness and diversity increase with stream size (Horwitz 1978; Vannote et al. 1980; Wiley et al. 1990), and other assemblage and population attributes, such as fish density, biomass, growth, body size, and trophic dynamics, change longitudinally (Matthews 1998). Thus, stream size or longitudinal position must be taken into account when comparing fish assemblage data among sites within and among drainage networks.

Typically, stream longitudinal position, which is generally related to channel size, is defined by stream order (Horton 1945; Strahler 1957) or watershed drainage area (in acres or hectares). Because of inconsistencies in definition and reduced precision associated with the discrete scale (consecutive integers) of stream order, drainage area is generally a more useful descriptor of stream size and position.

Fausch et al. (1984) presented a simple graphical technique to demonstrate the effect of stream longitudinal position on fish species richness and to approximate expected criteria values for IBI applications. When species richness is plotted against stream order or watershed area (\log_{10} transformed), the distribution

of sites forms a right triangle, where the hypotenuse forms a positive-sloped line of maximum species richness for a river system or region. The line of maximum species richness is used in IBI practice to define “excellent” species richness (metric score = 5), which varies with stream size; similarly, sites with richness falling below the maximum expected may be rated depending on the degree of deviation below the line for a given stream order or watershed area. The line of maximum expected values can be quantitatively derived by calculating the 95th-percentile regression (Blackburn et al. 1992) rather than by visually fitting a line to perceived maximum values. This technique can be applied to other IBI metrics associated with species composition (Table 15.2) and may be less relevant to metrics associated with trophic composition or fish abundance and condition, which appear to vary less with stream position or size (Karr et al. 1986).

Examination or adjustment of other fish assemblage structural indices, in addition to IBI metrics, should be considered in most site comparisons within and among drainage networks. In the absence of data to develop such relationships for a river system or a region, site comparisons relevant to environmental quality should be conducted among stream sites of similar longitudinal position or size. Furthermore, such spatial and size effects apply to limnetic zones of lakes and reservoirs and should be considered in analogous lentic comparisons among sites.

Reference systems. Reference systems to represent undisturbed or least-disturbed ecological conditions are a critical component of any biotic or ecological assessment (National Research Council 1992; Hughes 1995; Karr and Chu 1999). Such systems are critical as a benchmark for comparison to detect and understand effects of human activities on ecosystems and to serve as a goal for ecological restoration. Therefore, selection of a reference system or definition of reference conditions is of utmost importance in biotic assessment, but no clear criteria exist for such decisions.

Variation over space and time is key to identifying reference systems or conditions. A system or location within a region with highest IBI metric scores or other appropriate biotic or physical criteria (e.g., watershed land use or riparian disturbance) may serve as reference conditions. Similarly, information from the past, recent or historical, may provide qualitative or quantitative descriptions of predisturbance conditions. Ichthyological references, graduate theses, and state and federal agency reports can be valuable sources of historical data on fish assemblages for specific regions.

Searches for information on predisturbance conditions may be difficult but perhaps the only option in regions exposed to large-scale degradation. For example, only 42 high-quality, free-flowing rivers remain in the conterminous United States, and only 2% of the rivers in those states have features sufficient to receive federal protection (Benke 1990). Further, Hynes (1970) purported that it would be extremely difficult to find any stream that has not been altered by humans and impossible to find any such river—and that assessment was made more than 30 years ago. Typically, a system or site of least disturbance must be substituted for an undisturbed reference or reference conditions from another region may be

cautiously applied. Hughes (1995) examined common, alternative, and combined approaches for determining regional reference conditions.

15.3.2.5. *Biotic Integrity Indices in Practice*

The IBI was an important advance in biotic assessment methods that has evolved to a concept rather than a restrictive protocol. Fisheries and aquatic scientists are free to apply their own experience, perspective, and creativity into assessing ecological integrity based on fish and invertebrate assemblages, as well as physical attributes of aquatic habitats, riparian zones, and landscapes. Development of an IBI is a rather intensive endeavor that requires sampling a substantial number of sites and careful deliberation regarding metric and criteria development and refinement. However, there is no reason that investigators undertaking more limited assessments should not use individual IBI metrics or related assemblage structural indices (e.g., richness, diversity, or evenness) singly or in aggregate as quantitative assemblage characteristics for comparison and assessment.

As with other techniques in this chapter, several restrictions apply. For any biotic index to be meaningful, it must be based on a thorough and representative sample of the fish assemblage. Thus, sampling considerations (section 15.2) are an important aspect of any assessment program. It is essential that samples reflect the fish assemblage resulting from the physical and biotic environment rather than from a sampling bias that may vary among sites. As with most indices presented in this chapter, biotic indices are relative rather than absolute values, and their application should reflect that limitation. Finally, a biotic index must have a demonstrable empirical relationship to environmental quality to be meaningful, and such steps should be incorporated into biotic index development.

■ 15.4 METHODS TO COMPARE COMMUNITIES

Community indices are useful for summarizing and describing the structure of a fish assemblage. However, they cannot be used to compare the composition and relative abundance of species in two or more assemblages directly. For example, species richness does not take species identity into account, and the IBI does not provide an explicit means to determine how two or more assemblages differ. Several approaches have been developed by ecologists for directly comparing two or more communities, and most have been used to study fish assemblages. These techniques can be roughly categorized as (1) resemblance measures for quantifying the similarity among assemblages, (2) classification methods for grouping assemblages based on their structure, (3) ordination methods for examining the relationships among assemblages, (4) categorical data analysis methods for estimating the differences among assemblages, and (5) graphical techniques for displaying the relationships among assemblages. Each of the techniques discussed below has associated assumptions and limitations that can affect the validity of community comparisons. Consequently, ecologists often employ two or more techniques to maximize insight and validate patterns indicated by a single method

(Green and Vascotto 1978; Gauch 1982; Romesburg 1990). However, fisheries scientists should refrain from data dredging, that is, conducting several analyses until one produces results that appear to make the most sense or are statistically significant. Such an approach to community level analysis is fraught with problems and should be avoided (see Rexstad et al. 1988). Rather, fisheries scientists should carefully consider the objectives of their study and the system being examined and develop a set of questions to be addressed through their analyses.

15.4.1 Measures of Community Similarity and Their Characteristics

Fisheries biologists are often interested in quantifying the similarity among fish communities based on their species composition and abundance. Resemblance coefficients are used to measure the similarity (or dissimilarity) of two or more communities with one or more characteristics, such as species presence, abundance, density, or other community functional parameter (Gauch 1982; Romesburg 1990). Thus, resemblance coefficients are a useful and relatively flexible means for quantifying the similarity among fish assemblages. Similarity and dissimilarity, however, are generally descriptive measures rather than statistical estimates (correlation coefficients are an exception), and, as such, it is difficult to estimate the statistical significance of relationships. Such associated significance tests usually require the use of computer-intensive resampling techniques and specialized software and are of limited scientific value (e.g., Van Sickle 1997; Johnson 1999); hence, we do not recommend their general use.

There are several types of resemblance coefficients, and the use of each depends upon the characteristics of the data and study objectives. Binary coefficients are used to measure the similarity between two assemblages using only species' presence and absence. Ordinal coefficients are used when species (relative) abundances have been transformed into ranks presumably to minimize the influence of sampling bias. Quantitative coefficients require an estimate of species-specific abundance, such as density, relative abundance, and counts of the number of individuals. Although resemblance measures can differ markedly in their data requirements and calculation, the best and most useful share two desirable characteristics. First, resemblance measures should be independent of the number of individuals in a sample and the number of species in an assemblage. Second, they should increase regularly from a set minimum to a set maximum as two fish assemblages become increasingly similar (Wolda 1981). Another characteristic to consider is the sensitivity of a particular resemblance measure to size (abundance) displacements. Measures that are insensitive to size displacements consider two assemblages to be similar if their attributes (e.g., density of a species) differ by an additive or multiplicative function (Figure 15.4). Consequently, different resemblance measures can provide markedly different estimates of the similarity (or dissimilarity) between two assemblages. Fisheries scientists should be aware of these characteristics and use the resemblance measure that best meets their needs.

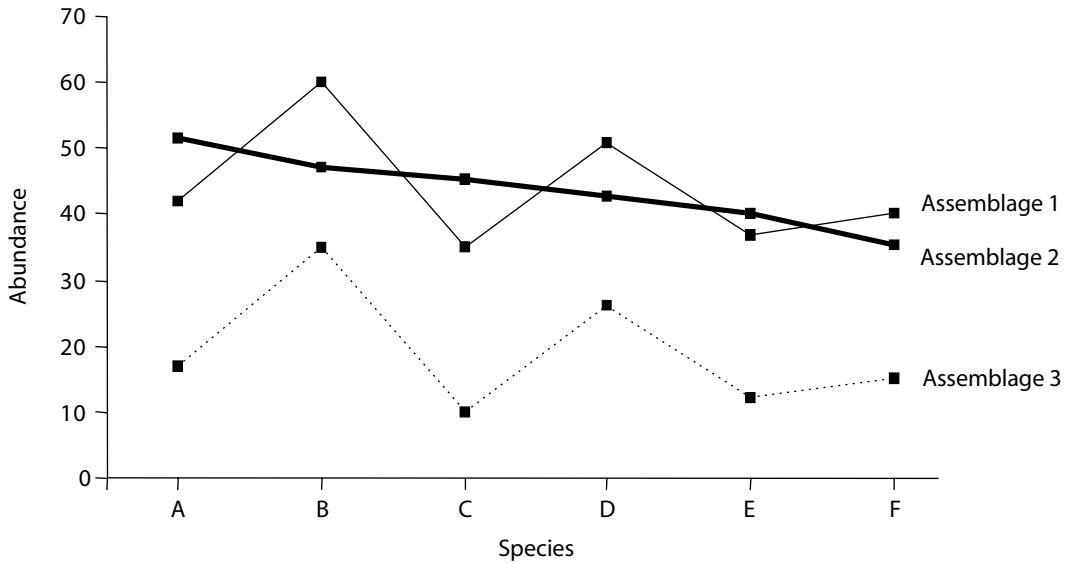


Figure 15.4 Species-specific abundances for three hypothetical fish assemblages. A resemblance measure that is insensitive to size displacement would score assemblages 1 and 3 as most similar because the species abundances differ by an additive constant (25), whereas a measure that is sensitive to size displacements would score assemblages 1 and 2 as most similar.

15.4.1.1 Coefficients for Binary (Nominal) Data

Species' presence and absence require a similarity coefficient that can be used with binary data (i.e., data that consist of two states). Binary coefficients are the simplest, but most imprecise, estimators of community similarity because they consider only species' presence or absence. Rare species and abundant species are weighted equally. Therefore, binary coefficients should be used only when species' presence are the only data available or in situations in which sampling conditions prevented the estimation of species relative abundances.

The best and most commonly used binary coefficients are the Jaccard's and simple matching coefficients, which vary from 0 to 1, with 0 indicating no species in common and 1 indicating identical species composition. Both measures are independent of the number of individuals in an assemblage (sample) and are insensitive to size displacements.

Jaccard's coefficient. The similarity, C , between a pair of assemblages j and k is calculated as

$$C_{jk} = \frac{p}{p + m}, \quad (15.8)$$

where p is the number of species that are present in both assemblages and m is the number of species present in one assemblage but not the other. An example calculation of Jaccard's coefficient is presented in Box 15.4.

Box 15.4 Calculation of Jaccard's and Simple Matching Coefficients

To calculate both Jaccard's and simple matching coefficients, first determine the number of species present and absent for both stations and the number of species occurring at one station but not another. Using the summary data for stations 1 and 2 on the Kankakee River, Illinois (Box 15.1),

number of species present at both stations is $p = 18$,
 number of species absent at both stations is $a = 4$, and
 number of species present at one station but not the other is $m = 12$.

Jaccard's Coefficient

From equation (15.8),

$$C_{jk} = \frac{p}{p+m}, \text{ and}$$

$$C_{1,2} = \frac{18}{18+12} = 0.60.$$

Simple Matching Coefficient

From equation (15.9),

$$C_{jk} = \frac{p+a}{p+m+a}, \text{ and}$$

$$C_{1,2} = \frac{18+4}{18+12+4} = 0.65.$$

Program

The following SAS program uses the DISTANCE macro, included in SAS/STAT software (version 6.09 or later; SAS Institute 2004), to compute the Jaccard and simple matching similarity coefficients for fish assemblages at all stations of the Kankakee River example data. Note that the path following the %INC must be changed to that of the SAS folder containing XMACRO, STDIZE, and DISTNEW macros on your computer or network. These macros are included with all versions of SAS/STAT software.

```

OPTIONS PS = 60 LS=78;
DATA SPECIES;
INPUT STATION $ LOG GZS BLM BUM CAP HOC MIS RDS RYS SAS SFS STS SUM BLR GOR NHS
SHR QLL RVR SVR SAB BKS BLG GSF LMB LOS OSF ROB SMB BAD BLD JOD LOP SLD;
LINES;
STATION1 1 1 1 0 1 0 1 0 1 1 1 1 0 0 1 1 1 1 1 0 1 1 1 0 1 1 1 1 1 1 1 0 1 1
STATION2 1 1 1 0 1 0 1 1 1 1 1 1 0 0 0 1 0 1 0 1 1 1 0 1 1 1 0 1 1 0 1 1 0 1 1 0 0

```

```

STATION3 0 1 1 0 1 0 1 1 1 1 1 1 0 1 1 1 1 1 1 0 1 0 1 1 1 0 1 1 0 1 1 1 1
STATION4 1 1 1 1 1 0 0 0 1 1 1 1 0 0 1 1 1 1 0 1 0 1 1 1 0 1 1 1 1 1 0 0 1 1
STATION5 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1
STATION6 1 1 1 0 1 0 0 1 1 1 1 1 1 0 1 1 1 1 1 1 0 1 0 1 0 1 1 1 1 1 1 0 0 1 1
;
%INC '<location of SAS sample folder>/XMACRO.SAS';
%INC '<location of SAS sample folder>/STDIZE.SAS';
%INC '<location of SAS sample folder>/DISTNEW.SAS';
%DISTANCE(DATA=SPECIES, OPTIONS=NOMISS, SHAPE = SQUARE, ID = STATION,
OUT=JACCARD, METHOD = JACCARD);
PROC PRINT;
%DISTANCE(DATA=SPECIES, OPTIONS=NOMISS, SHAPE = SQUARE, ID = STATION, OUT=MATCH,
METHOD = MATCH);
PROC PRINT;
RUN;

```

Program Output

Table Jaccard's coefficient and simple matching coefficient similarity matrices for fish assemblages at all stations of the Kankakee River example data (Box 15.1).

Station	Station					
	1	2	3	4	5	6
Jaccard Similarity Matrix						
1	1.0000	0.6000	0.7333	0.8519	0.7273	0.8214
2	0.6000	1.0000	0.7143	0.5333	0.6563	0.5667
3	0.7333	0.7143	1.0000	0.6129	0.7273	0.7000
4	0.8519	0.5333	0.6129	1.0000	0.7188	0.8148
5	0.7273	0.6563	0.7273	0.7188	1.0000	0.6471
6	0.8214	0.5667	0.7000	0.8148	0.6471	1.0000
Simple Matching Similarity Matrix						
1	1.0000	0.6471	0.7647	0.8824	0.7353	0.8529
2	0.6471	1.0000	0.7647	0.5882	0.6765	0.6176
3	0.7647	0.7647	1.0000	0.6471	0.7353	0.7353
4	0.8824	0.5882	0.6471	1.0000	0.7353	0.8529
5	0.7353	0.6765	0.7353	0.7353	1.0000	0.6471
6	0.8529	0.6176	0.7353	0.8529	0.6471	1.0000

The Jaccard coefficient is sensitive to the direction of the coding (i.e., asymmetric); hence, presence or absence should be coded the same for all species. This measure considers only mutual presence of a species, which should minimize the influence of false absences (i.e., a species is missed) in sampling data. Therefore, Jaccard's is the preferred method for analyzing fish assemblage similarity based on species presence.

Simple matching coefficient. The similarity, C , between j and k is calculated as

$$C_{jk} = \frac{p + a}{p + m + a}, \quad (15.9)$$

where p and m are defined above and a is the number of species absent in both assemblages. An example calculation of the simple matching coefficient is presented in Box 15.4.

In contrast to Jaccard's, the simple matching coefficient uses the mutual absence of a species and should be used only when there is no potential for false absences (i.e., missed species) in the data. Use of the simple matching coefficient requires an exact definition of the species pool, which is subjective to some degree, and large species pools may inflate the number of mutual absences and, thus, similarity.

15.4.1.2 Coefficients for Ranked (Ordinal) Data

The most widely used coefficients for ranked data are also nonparametric correlation coefficients. These resemblance measures are estimated using ranks in place of actual abundance estimates, and as such, they are unaffected by nonlinear relationships between species abundances of two assemblages. Ordinal measures are not as crude as binary coefficients but are less sensitive than are quantitative coefficients (section 15.4.1.3). Hence, we recommend their use over quantitative measures only when species abundance estimates are believed to be poor due to factors such as sampling difficulties.

The most commonly used ordinal coefficients are Spearman's rank correlation and Kendall's tau (Romesburg 1990). In contrast to most resemblance measures, these coefficients have values that vary from -1 to 1 , with -1 indicating different species assemblages and 1 indicating identical species composition. Both measures are strongly affected by the total number of individuals in a sample, especially when there are large numbers of species, and are insensitive to size displacements (Krebs 1998). They are most useful for comparing assemblages in low-diversity communities but should never be used as similarity measures when more than half of the abundances in one or more assemblage samples are zero (Field 1970).

Spearman's rank correlation. To estimate Spearman's rank correlation coefficient, fish abundance data are first sorted and ranked ($1 = \text{lowest}$) for each station. Ties (i.e., species with equal abundances) are assigned the average of the ranks. For example, two species with equal abundances in the fifth and sixth positions in the

sorted list would receive the rank 5.5. The similarity, θ , between assemblages r and s is then calculated using the Pearson product-moment correlation as

$$\theta_{rs} = \frac{\sum (r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum (r_i - \bar{r})^2 \sum (s_i - \bar{s})^2}}, \quad (15.10)$$

where r_i and s_i are the ranks corresponding to species i , and \bar{r} and \bar{s} are the mean ranks of all species in assemblage r and s , respectively. An alternative formula for estimating the Spearman's rank correlation coefficient can be found in Daniel (1990). An example calculation of Spearman's rank correlation coefficient is presented in Box 15.5.

Kendall's tau. To estimate Kendall's tau coefficient for assemblages j and k , species abundances are sorted from smallest to largest for assemblage j . If ties are present in assemblage j (i.e., two or more species have the same abundance), species abundances for those species within assemblage k are sorted in increasing magnitude within each tied group. The similarity, τ , is then calculated as

$$\tau_{jk} = \frac{\sum P_i - Q_i}{n(n-1)(0.5)}, \quad (15.11)$$

where P_i and Q_i are the total number of species in assemblage k with higher ranks and abundances greater (P_i) and less than (Q_i) species i in assemblage j , and n is the total number of species.

If ties are present in assemblage j , corresponding species in assemblage k are not counted for P_i and Q_i within each tied group. The denominator in equation (15.11) is also adjusted for ties as

$$\tau_{jk} = \frac{\sum P_i - Q_i}{\sqrt{n(n-1)(0.5) - T_j} \sqrt{n(n-1)(0.5) - T_k}}, \quad (15.12)$$

where $T_j = (0.5)\sum t_j(t_j - 1)$; $T_k = (0.5)\sum t_k(t_k - 1)$; t_j and t_k are the number of species that are tied at a given rank in assemblage j and k , respectively; and n is the total number of species. An example calculation of Kendall's tau coefficient with ties is presented in Box 15.5.

15.4.1.3 Coefficients for Quantitative Data

The best means to characterize the similarity among fish assemblages is through the use of quantitative resemblance coefficients. In contrast to binary and ordinal measures, these coefficients use species abundance estimates and are, therefore, much more sensitive to small differences between two fish assemblages. There is a wide variety of similarity measures available for use with quantitative data. However, the two best and most widely used measures in ecology are the percent similarity

Box 15.5 Calculation of Spearman's Rank Correlation and Kendall's Tau Similarity Coefficients

Species total abundances from stations 1 and 2 of the Kankakee River, Illinois (Box 15.1), are used to illustrate the calculation of Spearman's rank and Kendall's tau coefficients.

Table Species sorted by abundance and ranked for stations 1 and 2 of Kankakee River example (Box 15.1). Also included are computations for Kendall's tau that tally the number of species at station 2 with abundances higher (P_i) or lower (Q_i) than a given species at station 1.

Species and summary statistics	Sorted abundance		Rank abundance		Number of species at station 2 with abundances	
	Station 1	Station 2	Station 1	Station 2	Higher (P_i)	Lower (Q_i)
Hornyhead chub	0	0	4.5	6.5	18	0
Suckermouth minnow	0	0	4.5	6.5	18	0
Bullhead minnow	0	0	4.5	6.5	18	0
Black redhorse	0	0	4.5	6.5	18	0
Smallmouth buffalo	0	3	4.5	17	14	11
Largemouth bass	0	3	4.5	17	14	11
Johnny darter	0	6	4.5	19	14	12
Redfin shiner	0	22	4.5	24.5	9	16
Orangespotted sunfish	1	0	10.5	6.5	15	0
Blackside darter	1	9	10.5	21	10	12
Sand shiner	1	22	10.5	24.5	8	14
Green sunfish	1	42	10.5	29	5	17
River redhorse	2	0	13.5	6.5	15	0
Bluegill	2	0	13.5	6.5	15	0
Banded darter	4	0	15.5	6.5	14	0
Brook silverside	4	10	15.5	22	9	9
Slenderhead darter	5	0	18	6.5	12	0
Quillback	5	2	18	14.5	11	4
Northern hog sucker	5	2	18	14.5	11	4
Longnose gar	6	7	20	20	9	5
Silver redhorse	8	1	21.5	13	9	3
Rosyface shiner	8	89	21.5	32	2	10
Mimic shiner	10	11	23	23	7	4
Common carp	13	58	24	30	3	7
Spotfin shiner	19	24	25	26	5	4
Logperch	25	0	26	6.5	6	0
Rock bass	30	3	27	17	5	2
Golden redhorse	34	0	28	6.5	5	0
Shorthead redhorse	35	0	29.5	6.5	4	0
Longear sunfish	35	94	29.5	34	0	4
Bluntnose minnow	42	33	31	28	2	1
Striped shiner	45	32	32	27	2	0
Smallmouth bass	143	59	33	31	1	0
Gizzard shad	164	90	34	33	0	0
Average rank	17.5	17.5				
Sum			308	150		

Spearman's Rank Correlation Coefficient

To estimate Spearman's rank, species abundances are sorted in ascending order and assigned ranks (1 = lowest) for each assemblage. Ties are assigned the average of the tied ranks. The first eight species at station 1 have abundances of 0, and their average rank is calculated as

$$\frac{1 + 2 + 3 + \dots + 8}{8} = \frac{36}{8} = 4.5.$$

From equation (15.10),

$$\theta_{rs} = \frac{\sum (r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum (r_i - \bar{r})^2 \sum (s_i - \bar{s})^2}}, \text{ and}$$

$$\begin{aligned} \theta_{1,2} &= \frac{(4.5 - 17.5)(6.5 - 17.5) + (4.5 - 17.5)(6.5 - 17.5) + \dots + (34 - 17.5)(33 - 17.5)}{\sqrt{[(4.5 - 17.5)^2 + (4.5 - 17.5)^2 + \dots + (34 - 17.5)^2][(6.5 - 17.5)^2 + (6.5 - 17.5)^2 + \dots + (33 - 17.5)^2]}} \\ &= \frac{1308.250}{\sqrt{(3221.500)(3126.500)}} \\ &= 0.412. \end{aligned}$$

Kendall's Tau

To estimate Kendall's tau, species abundances are sorted for station 1 from lowest to highest, and because ties occur at station 1, species abundances for station 2 are also sorted within each tied group. For example, the first eight species at station 1 of the Kankakee River, hornyhead chub through redbfin shiner, have abundances of 0; hence, they are also sorted by their station 2 abundances as shown in the table above.

From equation (15.12),

$$\tau_{jk} = \frac{\sum P_i - Q_i}{\sqrt{n(n-1)(0.5) - T_j} \sqrt{n(n-1)(0.5) - T_k}}.$$

Using the sorted abundances for station 2, P_i is estimated for each species (i) by counting the number of other species at station 2 (see table) with greater abundance. For hornyhead chub, this is any species with abundance greater than 0. However, four of these species, smallmouth buffalo through redbfin shiner, have abundances greater than 0 but are not counted because their abundances are tied (all 0) at station 1. This leaves a total of 18 species with greater abundance than hornyhead chub at station 2. Then Q_i is estimated similarly by counting the number of species with lower abundance, which for hornyhead chub is none (0). The adjustment for ties in the denominator of equation (15.12), T_j and T_k , is estimated for each station by first counting the number of

(Box continues)

Box 15.5 (continued)

species for each tied abundance value. At station 1, eight species had 0 abundance, 4 had abundances of 1, 2 had abundances of 2, and so on. The adjustments are then estimated as

$$T_k = (0.5) \sum t_k(t_k - 1)$$

$$T_1 = (0.5)[(8)(8 - 1) + (4)(4 - 1) + (2)(2 - 1) + (2)(2 - 1) + (3)(3 - 1) + (2)(2 - 1) + (2)(2 - 1)] \\ = 41, \text{ and}$$

$$T_2 = (0.5)[(12)(12 - 1) + (2)(2 - 1) + (3)(3 - 1) + (2)(2 - 1)] \\ = 71.$$

Thus, replacing the symbols in equation (15.12) with their corresponding values

$$\tau_{1,2} = \frac{308 - 150}{\sqrt{34(34 - 1)(0.5) - 41} \sqrt{34(34 - 1)(0.5) - 71}} \\ = 0.313.$$

Program

The following SAS program computes Spearman's rank and Kendall's tau similarity coefficients for fish assemblages at stations on the Kankakee River.

```

OPTIONS PS = 60 LS=78;
DATA SPECIES;
INPUT SPECIES $ STATION1 STATION2 STATION3 STATION4 STATION5 STATION6;
LINES;
Longnose_gar 6 7 0 4 26 5
(input remaining 33 species)
;

```

and Morisita's indices. Percent similarity is used on relative species abundance data, and Morisita's index is employed when data consist of the number of individuals (whole numbers) per species.

Percent similarity. To calculate the percent similarity index, species abundances in each assemblage must first be standardized to percentages by dividing the abundance of each species in a sample by the total number of fish in the sample and multiplying by 100. The similarity, P , between assemblages j and k is calculated as

$$P_{jk} = \sum \text{minimum}(p_{ki}, p_{ji}), \quad (15.13)$$

where p_{ji} and p_{ki} are the relative abundances of species i in assemblage j and k , respectively, and minimum indicates that the smallest of the two relative abundances is used in the summation. An example calculation of the percent similarity index is presented in Box 15.6.

```

PROC CORR NOPRINT OUTS=SPEARMAN OUTK=KENDALL;
DATA SPEARMAN; SET SPEARMAN; WHERE _TYPE_='CORR'; DROP _TYPE_;
PROC PRINT;
DATA KENDALL; SET KENDALL; WHERE _TYPE_='CORR'; DROP _TYPE_;
PROC PRINT;
RUN;

```

Program Output

Table Spearman's rank and Kendall's tau similarity matrices for fish assemblages at stations 1 and 2 of the Kankakee River data (Box 15.1).

Station	Station					
	1	2	3	4	5	6
Jaccard Similarity Matrix						
1	1.0000	0.4122	0.7087	0.7318	0.6943	0.8540
2	0.4122	1.0000	0.4532	0.2912	0.4378	0.4087
3	0.7087	0.4532	1.0000	0.6372	0.4394	0.6565
4	0.7318	0.2912	0.6372	1.0000	0.6601	0.7953
5	0.6943	0.4378	0.4394	0.6601	1.0000	0.6989
6	0.8540	0.4087	0.6565	0.7953	0.6989	1.0000
Kendall's Tau Similarity Matrix						
1	1.0000	0.3130	0.5448	0.5762	0.5470	0.7060
2	0.3130	1.0000	0.3609	0.2374	0.3261	0.3288
3	0.5448	0.3609	1.0000	0.4966	0.3169	0.4898
4	0.5762	0.2374	0.4966	1.0000	0.4890	0.6385
5	0.5470	0.3261	0.3169	0.4890	1.0000	0.5106
6	0.7060	0.3288	0.4898	0.6385	0.5106	1.0000

The percent similarity index, also known as the Renkonen index after its creator (Renkonen 1938), is one of the best quantitative similarity measures (Wolda 1981). It varies from 0 to 100%, with 0 indicating no species in common and 100% indicating identical species composition. It is a very robust measure that is not influenced by the number of individuals in a sample and is insensitive to size displacements.

Morisita's index. The similarity, C , between assemblages j and k is calculated following Morisita (1959) as

$$C_{jk} = \frac{2 \sum X_{ij} X_{ik}}{(\lambda_j + \lambda_k) N_j N_k}, \quad (15.14)$$

where

$$\lambda_j = \frac{\sum [X_{ij}(X_{ij} - 1)]}{N_j(N_j - 1)}, \text{ and } \lambda_k = \frac{\sum [X_{ik}(X_{ik} - 1)]}{N_k(N_k - 1)}.$$

The number of individuals of species i is given by X_{ij} and X_{ik} , and N_j and N_k are the total number of individuals in assemblage j and k , respectively. Horn (1966) developed a simplified version of the index in which each λ is calculated without subtracting 1 from the total number of individuals in the assemblage. This modified version of Morisita's index is used only when abundance is expressed as a proportion, such as relative abundance and density. An example calculation of the Morisita's index is given in Box 15.6.

Morisita's index varies from 0 to 1, with 0 indicating no species in common, and 1 indicating identical species composition. Unlike other similarity coefficients, Morisita's index can be interpreted as a probability (Krebs 1998). It is not significantly influenced by the number of individuals in an assemblage sample, unless that total is very small, and is insensitive to size displacements (Wolda 1981).

15.4.1.4 Distance Measures

Distance coefficients are a special kind of quantitative resemblance measure that are used to estimate the dissimilarity between two fish assemblages. In contrast to similarity, low distance (dissimilarity) values indicate that two assemblages are more similar to one another with 0 indicating identical species composition. Interestingly, any of the similarity measures presented in this chapter can be transformed into a dissimilarity measure by multiplying its value by -1 or by subtracting from a constant corresponding to the maximum value (e.g., subtract from 100 the percent similarity index). Distance coefficients are generally used in cluster analysis (section 15.4.3) and require a quantitative measure of species-specific abundance, such as numbers of individuals, relative abundance, and density.

The most commonly used distance measures are the Euclidean and Bray–Curtis coefficients. Both measures are strongly affected by the total number of individuals and number of species in a sample and are the only resemblance measures presented in this chapter that are sensitive to size displacements.

Euclidean distance. The Euclidean distance, d , between assemblages j and k is calculated as

$$d_{jk} = \sqrt{\sum (X_{ij} - X_{ik})^2}, \quad (15.15)$$

where X_{ij} and X_{ik} are the abundances of species i in assemblage j and k , respectively. Euclidean distance values are strongly influenced by the number of species in an assemblage. To minimize this effect, researchers often calculate the average Euclidean distance, d' , as

$$d'_{jk} = \sqrt{\frac{\sum (X_{ij} - X_{ik})^2}{n}}, \quad (15.16)$$

where n is the total number of species. Example calculations of Euclidean distance and average Euclidean distance are presented in Box 15.7. Euclidean distance and average Euclidean distance can both vary from 0 to infinity, with 0

Box 15.6 Calculation of Percent Similarity and Morisita's Indices

Below we calculate percent similarity and Morisita's similarity indices based on summary fish abundance data from stations 1 and 2 on the Kankakee River, Illinois (Box 15.1).

Percent Similarity

To calculate percent similarity, species abundances must be expressed as percentages. Totals of 648 and 622 fish were collected from stations 1 and 2, respectively. Thus, species-specific abundances at each station are divided by their corresponding station totals and multiplied by 100.

From equation (15.13),

$$P_{jk} = \sum \text{minimum} (p_{kij}, p_{ji}), \text{ and}$$

$$\begin{aligned} P_{1,2} &= \text{minimum} (0.926, 1.125) + \text{minimum} (25.309, 14.469) + \dots + \text{minimum} (0.772, 0.000) \\ &\quad \text{(longnose gar)} \qquad \qquad \text{(gizzard shad)} \qquad \qquad \text{(slenderhead darter)} \\ &= 0.926 + 14.469 + \dots + 0.000 \\ &= 50.185\%. \end{aligned}$$

Morisita's Similarity Index

From equation (15.14),

$$C_{jk} = \frac{2 \sum X_{ij} X_{ik}}{(\lambda_j + \lambda_k) N_j N_k}.$$

$$\begin{aligned} \lambda_1 &= \frac{6(6-1) + 164(164-1) + \dots + 5(5-1)}{648(648-1)} \\ &= 0.135, \text{ and} \end{aligned}$$

$$\begin{aligned} \lambda_2 &= \frac{7(7-1) + 90(90-1) + \dots + 0(0-1)}{622(622-1)} \\ &= 0.096. \end{aligned}$$

$$\begin{aligned} C_{1,2} &= \frac{2[(6)(7) + (164)(90) + \dots + (5)(0)]}{(\lambda_j + \lambda_k)(648)(622)} \\ &= \frac{63,236.000}{92,875.079} \\ &= 0.681. \end{aligned}$$

indicating identical assemblages and large distances indicating very different assemblage structure.

Bray-Curtis coefficient. The distance, b , between assemblage j and k is calculated following Bray and Curtis (1957) as

$$b_{jk} = \frac{\sum |X_{ij} - X_{ik}|}{\sum (X_{ij} + X_{ik})}, \quad (15.17)$$

where X_{ij} and X_{ik} are the abundance of species i in assemblage j and k , respectively. An example calculation of the Bray–Curtis coefficient is presented in Box 15.7. The Bray–Curtis coefficient varies from 0 to 1, with 0 indicating identical assemblages and 1 indicating no species in common. It also tends to be strongly influenced by abundant species and should not be used when fish assemblage samples are dominated by a few very abundant species (i.e., evenness is low, Wolda 1981).

Box 15.7 Calculation of Euclidean and Bray–Curtis Distances.

Euclidean and Bray–Curtis distances are calculated for summary fish abundance data from stations 1 and 2 of the Kankakee River, Illinois (Box 15.1).

Euclidean Distance

From equation (15.15),

$$d_{jk} = \sqrt{\sum (X_{ij} - X_{ik})^2}.$$

$$d_{1,2} = \sqrt{(6 - 7)^2 + (164 - 90)^2 + \dots + (5 - 0)^2}$$

(longnose gar) + (gizzard shad) + ... + (slenderhead darter)

$$d_{1,2} = \sqrt{31,488.00} = 177.449.$$

Average Euclidean Distance

From equation (15.16),

$$d'_{jk} = \sqrt{\frac{\sum (X_{ij} - X_{ik})^2}{n}}.$$

$$d'_{1,2} = \sqrt{\frac{31,488.00}{34}} = 30.432.$$

Bray–Curtis coefficient

From equation (15.17),

$$b_{jk} = \frac{\sum |X_{ij} - X_{ik}|}{\sum (X_{ij} + X_{ik})}.$$

$$b_{1,2} = \frac{|6 - 7| + |164 - 90| + \dots + |5 - 0|}{(6 + 7) + (164 + 90) + \dots + (5 + 0)}$$

$$= \frac{630}{1,270} = 0.496.$$

15.4.2 Classification Techniques

Similarity coefficients are useful for examining relationships among small numbers of fish assemblages. Fisheries scientists, however, often need to compare several assemblages simultaneously for the purposes of grouping or classifying them based on their structure. Cluster analysis includes a set of techniques that can be used to examine the relationships among two or more communities and group

Program

The following SAS program uses the DISTANCE macro, included in SAS/STAT software (SAS Institute 2004), version 6.09 or later, to compute Euclidean distance for fish assemblages at all stations on the Kankakee River. Note that the path following the %INC must be changed to that of the SAS folder containing XMACRO, STDIZE, and DISTNEW macros on your computer or network.

```

OPTIONS PS = 60 LS=78;
DATA SPECIES;
INPUT SPECIES $ STATION1 STATION2 STATION3 STATION4 STATION5 STATION6;
LINES;
Longnose_gar 6 7 0 4 26 5
(input remaining 33 species)
;
PROC TRANSPOSE DATA = SPECIES OUT = SPECIES;
%INC '<location of SAS sample folder>/XMACRO.SAS';
%INC '<location of SAS sample folder>/STDIZE.SAS';
%INC '<location of SAS sample folder>/DISTNEW.SAS';
%DISTANCE(DATA=SPECIES, OPTIONS=NOMISS, SHAPE = SQUARE, ID = _NAME_,
OUT=EUCLID, METHOD = EUCLID);
PROC PRINT;
RUN;

```

Program Output

Table Euclidean distance matrix for fish assemblages at six stations on the Kankakee River (see Box 15.1).

Station	Station					
	1	2	3	4	5	6
1	0.0000	177.4486	175.2341	171.6945	275.2181	93.2202
2	177.4486	0.0000	218.1811	185.4104	380.0776	225.1000
3	175.2341	218.1811	0.0000	81.1788	435.1597	211.5349
4	171.6945	185.4104	81.1788	0.0000	435.5250	206.7970
5	275.2181	380.0776	435.1597	435.5250	0.0000	258.5672
6	93.2202	225.1000	211.5349	206.7970	258.5672	0.0000

them into classes (or clusters) based on their similarity (Romesburg 1990). Classification, however, is more of a skill than an exact science, and it requires a certain amount of ecological insight and knowledge of the systems being studied. There is no single best classification system for grouping fish assemblages or determining the exact number of groups to distinguish. Thus, a fish assemblage classification system developed for one purpose may be inappropriate for addressing another (Romesburg 1990). As with all analytical techniques, we encourage fisheries scientists to consider their objectives and planned application of their classification system carefully before attempting to develop fish assemblage classifications.

There are two basic types of cluster analyses—hierarchical and nonhierarchical—that have been used by biologists to develop polythetic classifications (i.e., classifications based on overall similarity). Among these, hierarchical methods are generally the simplest, most easy to use, and the only clustering methods considered here that are useful for examining relationships among assemblages. However, they can become cumbersome when the number of samples (i.e., assemblages) is large. Nonhierarchical methods are computationally complex and require the use of computer programs but are appropriate and useful when the number of samples is large. Below, we discuss the best, most widely used of these two clustering methods for ecological classification. For a more thorough treatment of cluster analysis and ecological classification, we recommend Romesburg (1990) and Everitt (1993).

15.4.2.1 *Hierarchical Cluster Analysis*

By far, the most widely used form of cluster analysis is hierarchical clustering. It is used to reveal relationships among assemblages based on resemblance measures (section 15.4.1). Hierarchical cluster analysis begins with a matrix of resemblance coefficients (see Box 15.6 for example). Pairs of assemblages are then grouped sequentially using a clustering method. Clustering begins by grouping the most similar (or least dissimilar) pair(s) of assemblages. The next most similar pair(s) is then clustered, and the process continues until all assemblages are contained in a single cluster. Results are displayed in a diagram, called a dendrogram or tree, that shows the similarities in the form of a hierarchy (hence, the name). Relationships indicated by hierarchical cluster analysis are significantly influenced by the characteristics of the resemblance measure and the clustering method. In fact, different combinations of resemblance measures and clustering methods can provide quite disparate estimates of relationships among assemblages (Figure 15.5). Because there is no truly objective means of clustering (Romesburg 1990; Krebs 1998), fisheries biologists should consider the characteristics of resemblance measures and clustering methods and choose those that make the most sense. Having previously detailed the characteristics of resemblance measures, we now describe those of clustering methods.

Single linkage. Single-linkage clustering is the simplest form of hierarchical cluster analysis. It begins with a resemblance matrix and uses the nearest-neighbor rule to define similarity or dissimilarity among clusters. For distance measures, assemblages are clustered as follows.

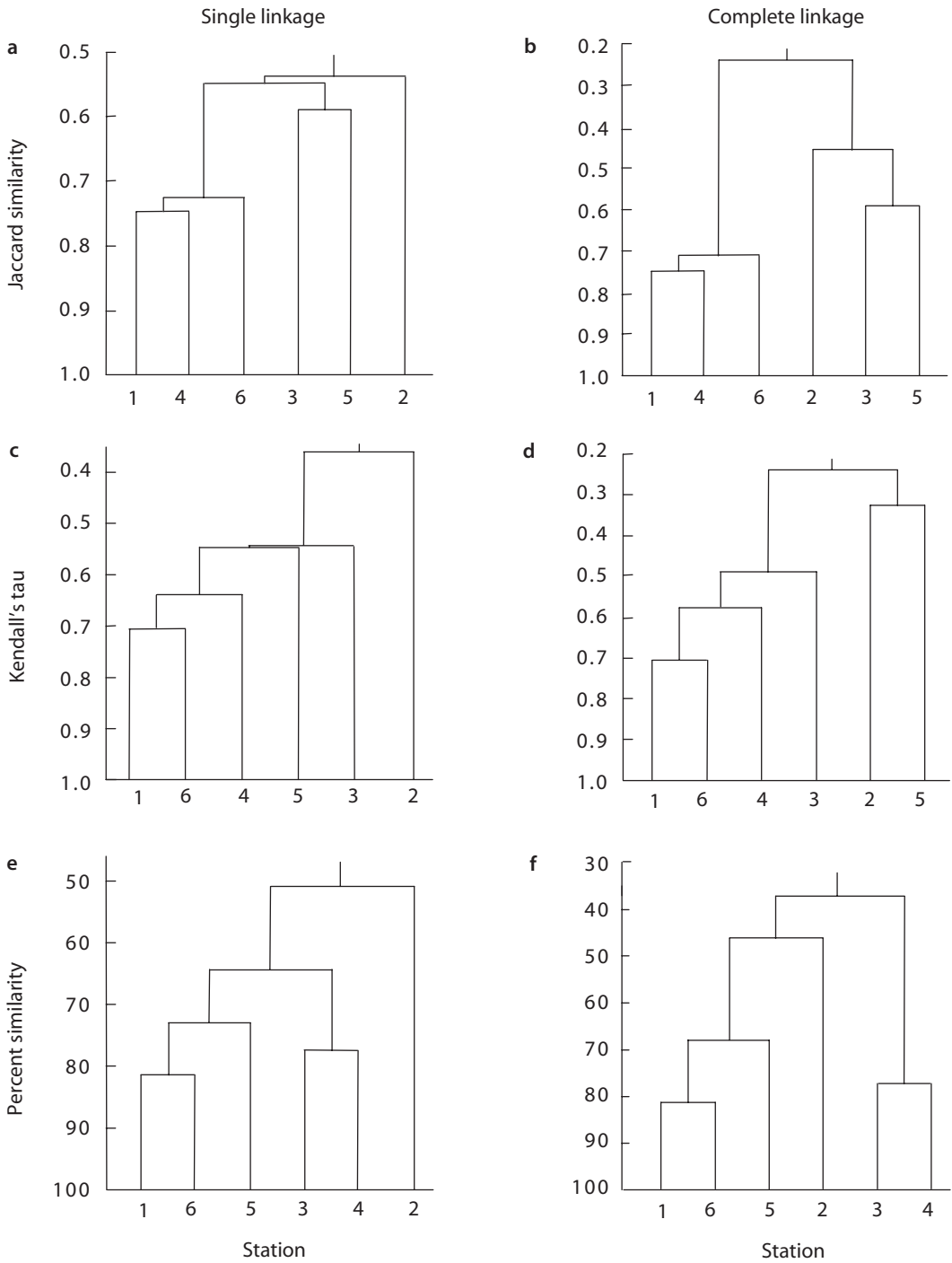


Figure 15.5 Cluster analyses of fish assemblages at six stations on the Kankakee River, Illinois (Box 15.1). Trees (a) and (b) were created using Jaccard's similarity coefficient, (c) and (d) with Kendall's tau, and (e) and (f) with the percent similarity index. Single-linkage clustering method was used for trees (a), (c), and (e) and complete-linkage for trees (b), (d), and (f).

Step 1. Find the pair of assemblages with the smallest distance (most similar) and combine into the first cluster.

Step 2. Find the second pair with the smallest distance and cluster. This pair can include two assemblages or an assemblage and the cluster created in step 1 whose nearest-neighbor distance, d , from an assemblage j is defined as

$$d_{j(m,n)} = \text{minimum} (d_{jm}, d_{jn}), \quad (15.18)$$

where (m, n) is a cluster containing assemblages m and n , and minimum indicates that the smallest distance is used in clustering.

Step 3. Find the third pair with the smallest distance and group (cluster). This pair can include two assemblages, an assemblage and a cluster, or two clusters whose nearest-neighbor distance, d , is defined as

$$d_{(j,k)(m,n)} = \text{minimum} (d_{jm}, d_{jn}, d_{km}, d_{kn}), \quad (15.19)$$

where (j, k) is the cluster containing assemblage j and k and (m, n) is defined above.

Step 3 is repeated until all of the assemblages are contained in one cluster. Single-linkage clustering is illustrated in Box 15.8.

For similarity measures, the nearest-neighbor distance is the highest similarity value, and hence, the minimum in equations (15.18) and (15.19) is replaced by a maximum.

Complete linkage. Complete-linkage clustering is also a relatively simple form of hierarchical clustering. It is very similar to single linkage except that the rules for defining distances among clusters are the exact opposite. Thus, the first step is identical to that described above for single linkage. Subsequent steps proceed as follows.

Step 2. Find the second pair with the smallest distance (most similar) and cluster. This pair can include two assemblages or an assemblage and the cluster created in step 1 whose farthest-neighbor distance, d , from an assemblage j is defined as

$$d_{j(m,n)} = \text{maximum} (d_{jm}, d_{jn}), \quad (15.20)$$

where (m, n) is a cluster containing assemblage m and n , and maximum indicates that the largest distance is used in clustering.

Step 3. Find the third pair with the smallest distance and cluster. Accordingly, the farthest-neighbor distance, d , between two clusters is defined as

$$d_{(j,k)(m,n)} = \text{maximum} (d_{jm}, d_{jn}, d_{km}, d_{kn}), \quad (15.21)$$

where (j, k) is a cluster containing assemblage j and k , and (m, n) is defined above. As above, step 3 is repeated until all of the assemblages are contained a single cluster.

Box 15.8 Clustering Procedure for Hierarchical Cluster Analysis

Below are dendrograms of fish assemblages at six stations on the Kankakee River, Illinois (Box 15.1), resulting from single- and average-linkage clustering of Euclidean distance resemblance measure of summary fish abundance data. Step labels correspond to clustering steps, detailed below.

Initial Resemblance Matrix (Euclidean Distance)

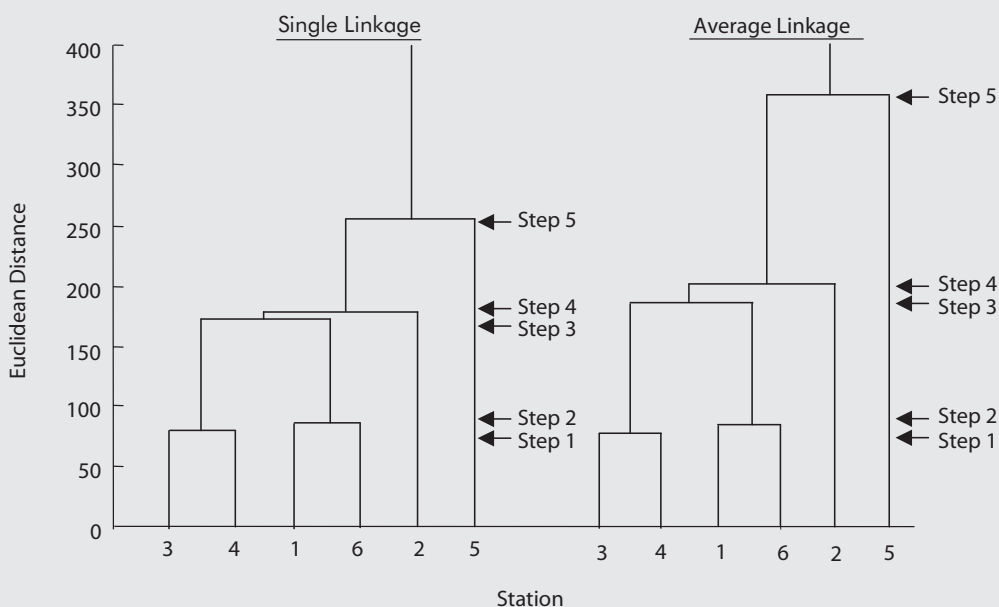


Table Euclidean distance matrix for fish assemblages at six stations of the Kankakee River data (Box 15.1).

Station	Station					
	1	2	3	4	5	6
1						
2	177.45					
3	175.23	218.18				
4	171.69	185.41	81.18			
5	275.22	380.08	435.16	435.53		
6	93.22	225.10	211.53	206.80	258.57	

Step 1

The first clustering step is to join the most similar pair of assemblages (i.e., smallest distance), which is stations 3 and 4, resulting in 1, 2, (3, 4), 5, and 6 at distance 81.18. The resemblance matrix is then

(Box continues)

Box 15.8 (continued)

adjusted to reflect the creation of (3, 4). For single-linkage clustering, the distance between (3, 4) and station 1 is estimated using equation (15.18), where

$$d_{j(m,n)} = \text{minimum} (d_{jm}, d_{jn}).$$

$$d_{1(3,4)} = \text{minimum} (175.23, 171.69)$$

$$= 171.69.$$

For average-linkage clustering, the distance between (3, 4) and station 1 is estimated using equation (15.22) where

$$d_{j(m,n)} = \frac{d_{jm} + d_{jn}}{N_{(m,n)}}$$

$$d_{1(3,4)} = \frac{175.23 + 171.69}{2}$$

$$= 173.46.$$

For each clustering method, compute the distances between (3,4) and the remaining stations and place into the corresponding revised matrix, shown below.

Table Second resemblance matrix based on single- and average-linkage clustering methods.

Station	Stations					Station	Stations				
	1	2	5	6	(3,4)		1	2	5	6	(3,4)
	Single-linkage distances						Average-linkage distances				
1						1					
2	177.45					2	177.45				
5	275.22	380.08				5	275.22	380.08			
6	93.22	225.10	258.57			6	93.22	225.10	258.57		
(3,4)	171.69	185.41	435.16	206.80		(3,4)	173.46	201.80	435.34	209.17	

Step 2

Next, join the most similar pair shown in the second resemblance matrix. For both clustering methods, this is stations 1 and 6, resulting in 2, (3,4), 5, and (1,6) at distance 93.22. As above, adjust each resemblance matrix to reflect the creation of (1,6). With single-linkage clustering, the distance between (1,6) and (3,4) is estimated using equation (15.19) and distances from the Euclidian distance matrix:

$$d_{(j,k)(m,n)} = \text{minimum} (d_{jm}, d_{jn}, d_{km}, d_{kn}).$$

$$d_{(3,4)(1,6)} = \text{minimum} (175.23, 171.69, 211.53, 206.80)$$

$$= 171.69.$$

With average-linkage clustering, the distance between (1,6) and (3,4) is estimated using equation (15.23) as

$$d_{(j,k)(m,n)} = \frac{d_{jm} + d_{jn} + d_{km} + d_{kn}}{N_{jk} + N_{mn}}$$

$$d_{(3,4)(1,6)} = \frac{175.23 + 171.69 + 211.53 + 206.80}{2 + 2}$$

$$= 191.31.$$

Compute the distances between (1,6) and the remaining stations with each clustering method, and place into the corresponding revised matrix, shown below.

Table Third resemblance matrix based on single- and average-linkage clustering methods.

Stations				Stations					
Station	2	5	(3,4)	(1,6)	Station	2	5	(3,4)	(1,6)
Single-linkage distances				Average-linkage distances					
2					2				
5	380.08				5	380.08			
(3,4)	185.41	435.16			(3,4)	201.80	435.34		
(1,6)	177.45	258.57	171.69		(1,6)	201.27	266.89	191.31	

Step 3

Join the most similar pair in the third resemblance matrix. For single-linkage distance, this is (3,4) and (1,6), resulting in 2, 5, and (1,3,4,6) at distance 171.69. With average-linkage distance, join (3,4) and (1,6) at 191.32. Compute the distances between (1,3,4,6) and stations 2 and 5 with each clustering method, and place into the corresponding revised matrix, shown below.

Table Fourth resemblance matrix based on single- and average-linkage clustering methods.

Stations				Stations			
Station	2	5	(1,3,4,6)	Station	2	5	(1,3,4,6)
Single-linkage distances				Average-linkage distances			
2				2			
5	380.08			5	380.08		
(1,3,4,6)	177.45	258.57		(1,3,4,6)	201.54	351.12	

Step 4

Join the most similar pair in the fourth resemblance matrix, 2 and (1,3,4,6), resulting in 5 and (1,2,3,4,6) at distance 177.45 for single linkage and 201.54 for complete linkage, respectively. Revise each resemblance matrix to reflect the creation of (1,2,3,4,6).

(Box continues)

Box 15.8 (continued)**Table** Fifth resemblance matrix based on single- and average-linkage clustering methods.

Station	Stations		Station	Stations	
	5	(1,2,3,4,6)		5	(1,2,3,4,6)
	Single-linkage distances		Average-linkage distances		
5			5		
(1,2,3,4,6)	258.57		(1,2,3,4,6)	356.91	

Step 5

Join 5 to (1,2,3,4,6) to create one cluster (1,2,3,4,5,6) at 258.567 and 356.91 for single and average linkage, respectively.

Program

The following SAS program clusters and plots a dendrogram of fish assemblages at six stations on the Kankakee River, Illinois, with single-linkage clustering method (METHOD = SINGLE) and Euclidean distance resemblance measure.

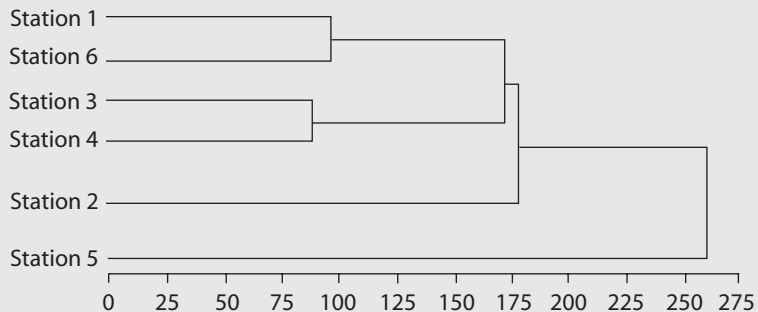
```

OPTIONS PS = 60 LS=78;
DATA EUCLID (TYPE = DISTANCE);
INPUT NAME $ STATION1 STATION2 STATION3 STATION4 STATION5 STATION6;
LINES;
STATION1 0.000 177.449 175.234 171.694 275.218 93.220
STATION2 177.449 0.000 218.181 185.410 380.078 225.100
STATION3 175.234 218.181 0.000 81.179 435.160 211.535
STATION4 171.694 185.410 81.179 0.0000 435.525 206.797
STATION5 275.218 380.078 435.160 435.525 0.0000 258.567
STATION6 93.220 225.100 211.535 206.797 258.567 0.000
;
PROC CLUSTER DATA = EUCLID OUTTREE = SINGTREE
METHOD = SINGLE NONORM NOSQUARE NOPRINT; ID NAME;
PROC TREE DATA = SINGTREE HORIZONTAL FC= 'C' DIS LEVEL = 0;
RUN;

```

Program Output

Name of observation or cluster

**Figure** Dendrogram of fish assemblages at six stations on the Kankakee River, Illinois, obtained with single-linkage clustering method and Euclidean distance resemblance measure.

For similarity measures, the farthest-neighbor distance is defined as the lowest similarity value of cluster members, and hence, the maximum in equations (15.20) and (15.21) is replaced by a minimum.

Average linkage. Average-linkage clustering, also known as unweighted pair-group with arithmetic averaging (Romesburg 1990), is the conceptual middle ground between single- and complete-linkage methods. The clustering steps are identical to those defined above. However, the distance between an assemblage, j , and a cluster (m, n) is defined as

$$d_{j(m,n)} = \frac{d_{jm} + d_{jn}}{N_{(m,n)}}, \quad (15.22)$$

and between two clusters as

$$d_{(j,k)(m,n)} = \frac{d_{jm} + d_{jn} + d_{km} + d_{kn}}{N_{jk} + N_{mn}}, \quad (15.23)$$

where N_{jk} and N_{mn} are the number of assemblages in clusters (j, k) and (m, n), respectively. Average-linkage clustering is illustrated in Box 15.8.

Dendrograms created with single-linkage clustering tend to be relatively long and narrow, whereas complete linkage tends to produce short, relatively compact trees with fewer clusters. Average linkage produces trees that are somewhat intermediate to these two extremes, and it is the most widely used clustering method. In a detailed analysis, Farris (1969) found that average linkage most faithfully represented the relationships among objects (e.g., assemblages) based on a mathematical analysis of different clustering methods. Nonetheless, we recommend that fisheries scientists create trees with two or more clustering methods and examine the fit of each. In the next section, we outline the method used to examine the adequacy of clustering methods.

Cophenetic correlation. The most appropriate means to assess the fit of each clustering method is by calculating a cophenetic correlation coefficient (Romesburg 1990). The cophenetic correlation coefficient is an unbiased measure of how well the cluster diagram represents relationships in the resemblance matrix; the largest correlation identifies the best clustering method. The cophenetic correlation coefficient is simply the Pearson product-moment correlation coefficient (equation [15.10]) between the resemblance matrix and the cophenetic matrix. The cophenetic matrix is an array of the distances among assemblages as represented in a dendrogram. It is estimated by tracing the path connecting each pair of assemblages in the dendrogram. An example calculation of the cophenetic correlation coefficient is presented in Box 15.9.

Classification. Classes are formed during hierarchical cluster analysis by cutting the tree at a specified level of similarity (Figure 15.6). Determining which level of similarity to cut the tree is highly subjective and depends upon study objectives (Romesburg 1990). For example, cutting a tree at a high level of similarity would

Box 15.9 Calculation of a Cophenetic Correlation Coefficient

Fish assemblages of the Kankakee River, Illinois (Box 15.1), were clustered with the single-linkage method (Box 15.8). Below, we illustrate calculation of the matrix cophenetic correlation coefficient. The values in the cophenetic matrix are estimated from the single-linkage dendrogram (Box 15.8) by tracing the path connecting each pair of assemblages. For example, when tracing the linkage from station 1 upward through the tree and downward to station 2, the greatest distance is 177.45. The remaining values are similarly estimated and are included in the cophenetic matrix below.

Table Resemblance matrix (Euclidean distance matrix) and cophenetic matrix from single-linkage clustering for calculation of cophenetic correlation coefficient.

Station	Station				
	1	2	3	4	5
Resemblance Matrix					
1					
2	177.45				
3	175.23	218.18			
4	171.69	185.41	81.18		
5	275.22	380.08	435.16	435.53	
6	93.22	225.10	211.53	206.80	258.57
Cophenetic Matrix					
1					
2	177.45				
3	171.69	177.45			
4	171.69	177.45	81.18		
5	258.57	258.57	258.57	258.57	
6	93.22	177.45	171.69	171.69	258.57

Table Side-by-side comparison of resemblance and cophenetic matrices for calculation of cophenetic correlation coefficient.

Matrix column, row	Resemblance matrix	Cophenetic matrix
1,2	177.45	177.45
1,3	175.23	171.69
1,4	171.69	171.69
1,5	275.22	258.57
1,6	93.22	93.22
2,3	218.18	177.45
2,4	185.41	177.45
2,5	380.08	258.57
2,6	225.10	177.45
3,4	81.18	81.18
3,5	435.16	258.57
3,6	211.53	171.69
4,5	435.53	258.57
4,6	206.80	171.69
5,6	258.57	258.57
Average	235.36	190.92

(Box continues)

Box 15.9 (continued)

The resemblance matrix then is unraveled and its values are paired with the corresponding values in the cophenetic matrix as shown above. The cophenetic correlation between the resemblance (r) and cophenetic (s) matrices is calculated using equation (15.10) as

$$\theta_{rs} = \frac{\sum (r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum (r_i - \bar{r})^2 \sum (s_i - \bar{s})^2}}$$

$$\begin{aligned} \theta_{rs} &= \frac{(177.45 - 235.36)(177.45 - 190.92) + \dots + (258.57 - 235.36)(258.57 - 190.92)}{\sqrt{[(177.45 - 235.36)^2 + \dots + (258.57 - 235.36)^2][(177.45 - 190.92)^2 + \dots + (258.57 - 190.92)^2]}} \\ &= \frac{77,129.30}{\sqrt{[162,332.91][46,673.49]}} \\ &= 0.886. \end{aligned}$$

Cophenetic correlation coefficients would be calculated for each clustering method. The hierarchical tree (i.e., dendrogram) with the largest cophenetic correlation would be selected to infer relationships among fish assemblages.

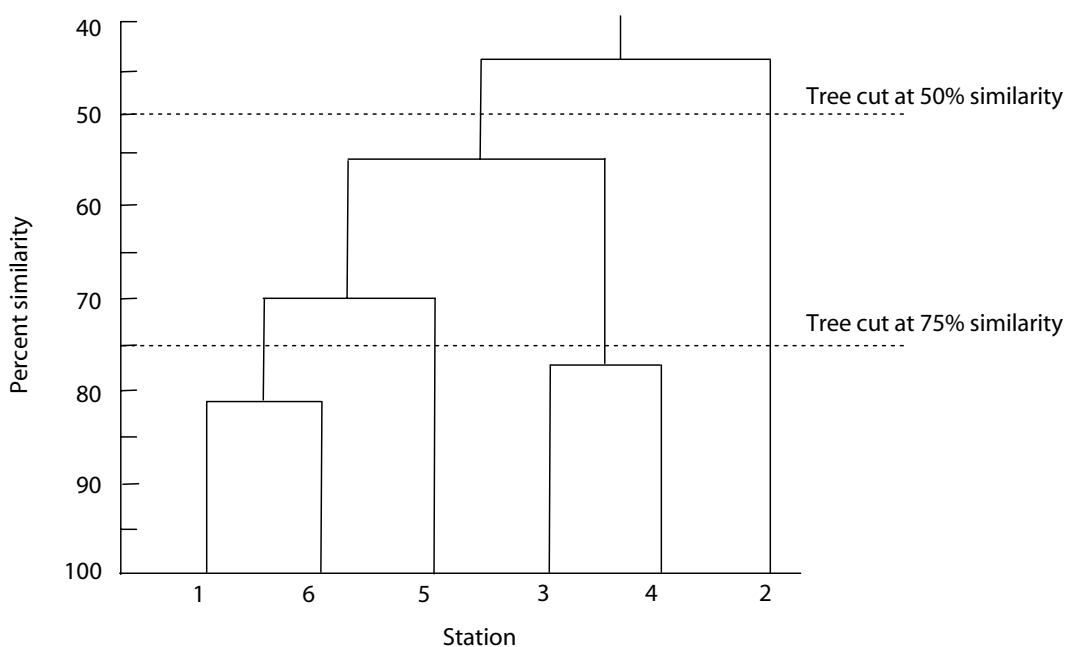


Figure 15.6 Dendrogram of fish assemblages at six stations on the Kankakee River, Illinois (Box 15.1), clustered using average linkage. Broken lines represent two cut points for classifying the fish assemblages. Four classes (1,6), 5, (3,4), and 2 are formed at 75% similarity, whereas two classes, (1,6,5,3,4) and 2, are formed at 50% similarity.

likely result in an excessively large number of groups, which would diminish one of the advantages of classification (i.e., reducing the data to a manageable size). Conversely, cutting a tree at a low level of similarity would likely introduce too much variation (heterogeneity) with each group, which could render the classification useless. Good classifications should attempt to minimize the number of groups created while simultaneously maximizing the within-group similarity.

15.4.2.2 *K-Means Cluster Analysis*

The most common form of nonhierarchical cluster analysis is *k*-means clustering. It is used to group assemblages into *k* clusters (i.e., *k* = number of clusters) based on their Euclidean distances. The *k*-means clustering procedures differ markedly from hierarchical clustering. It begins with a matrix of Euclidean distances and randomly assigns the assemblages to a prespecified number of clusters, *k*. Using a variance-minimizing algorithm, the assemblages are reassigned to different clusters on the basis of their similarity (distance) to other assemblages in a cluster. This process continues iteratively until the distance within each cluster is minimized and no assemblages need to be reassigned. The *k*-means clustering analysis is relatively robust to outlying data because, unlike hierarchical cluster analysis, the nature of the relationships among assemblages is unconstrained. However, the minimizing algorithm used for *k*-means clustering is inefficient when the number of samples is relatively low (Hartigan 1985). Therefore, we recommend *k*-means clustering only when the number of assemblages in the data set exceeds 30.

Similar to hierarchical cluster analysis, determining the number of clusters for a particular classification with *k*-means clustering is somewhat subjective. In fact, there are no completely unbiased methods for determining the number of clusters for any type of cluster analysis (Hartigan 1985). However for *k*-means clustering, statisticians have developed several methods that can be used to examine the relationship between within-cluster similarity and the number of clusters. This, in turn, can be used to select the optimal number of clusters (*k*) for a classification. One such method is to fit *k*-means clusters for several values of *k* and plot the overall R^2 versus the number of clusters. The overall R^2 is a measure of predictability of the fish assemblage within a cluster and is analogous to an r^2 in regression analysis. The optimal number of clusters (*k*) is considered the lowest value at which the R^2 begins to level off and reach an asymptote (Figure 15.7).

Another method for ascertaining the optimal number of clusters is the cubic clustering criterion (CCC) developed by Sarle (1983). Cubic clustering criterion is an estimator of *k*-means cluster fit. Similar to the overall R^2 method, the optimal number of clusters is determined by estimating CCC with *k*-means clustering for several values of *k* and examining a plot of *k* and CCC to find the smallest value of *k* with the lowest CCC (Figure 15.7). There are several other methods that have been developed for estimating the optimal number of clusters, all of which require specialized software and a strong background in statistical theory (e.g., Milligan and Cooper 1985; Mueller and Sawitzki 1991). For most fisheries applications, we suggest that scientists use both the R^2 and CCC to help find the optimal

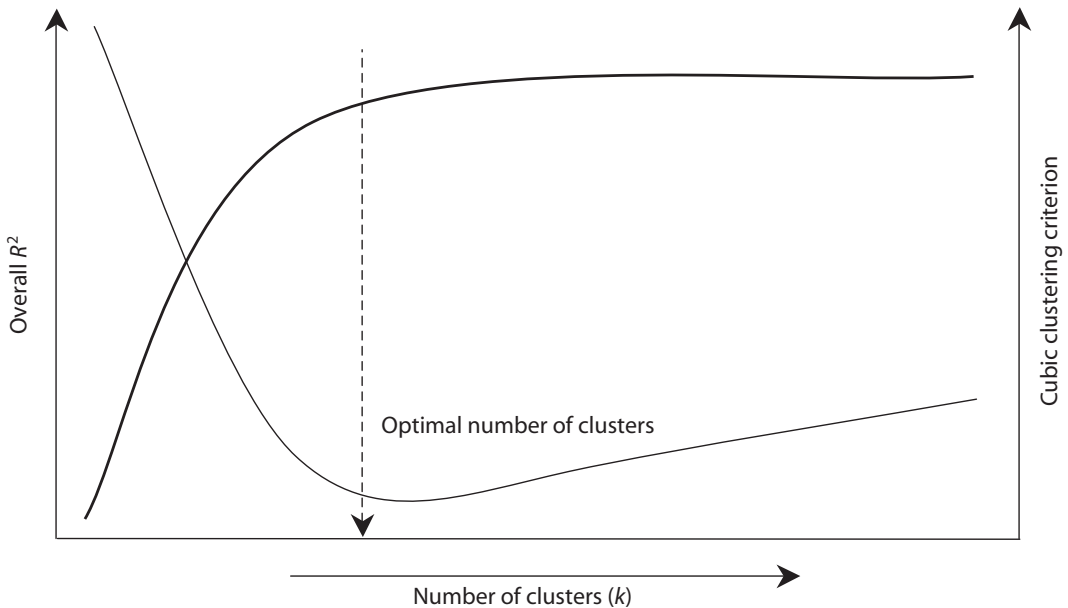


Figure 15.7 Hypothetical plot of the overall R^2 (thick line) and cubic clustering criterion (thin line) for various cluster sizes created using k -means clustering. Broken line represents the optimal number of clusters in the classification based on the overall R^2 and cubic clustering criterion.

number of clusters (k). An example SAS program for k -means clustering and the associated output are provided in Box 15.10.

15.4.3 Ordination Techniques

Ordination is the term for a variety of statistical techniques that ecologists have used to examine complex multivariate relationships among biotic communities and their environment. There are two broad classes of ordination—constrained and unconstrained. Constrained ordination is used to examine the relationship between assemblage structure and environmental gradients (e.g., stream habitat characteristics or water quality), whereas unconstrained ordination is used to examine the relationships among assemblages or species. Because we emphasize assemblage structure in this chapter, rather than function, we consider only unconstrained ordination. We refer fisheries scientists interested in learning more about constrained ordination to Ter Braak (1986) and Ter Braak and Prentice (1988).

Unconstrained ordination is used to summarize complex, multivariate relationships among assemblages and graphically display these within a small number of dimensions, usually two to three. Unlike cluster analysis, relationships are scored on a continuous scale (e.g., they need not be hierarchical or placed into discrete

Box 15.10 K-Means Clustering Analysis

The following SAS program performs *k*-means clustering with PROC FASTCLUS on the summary fish abundance data for six stations on the Kankakee River, Illinois (Box 15.1). Three-letter codes are used in place of species names (see Box 15.11 for key to codes). The number of clusters, $k = 3$, is specified by the MAXCLUSTERS command. Note that this example is for illustration only. The *k*-means clustering procedure should be used only when the number of assemblages (samples) exceeds 30.

Program

```

OPTIONS PS = 60 LS=78;
DATA SPECIES;
INPUT STATION $ LOG GZS BLM BUM CAP HOC MIS RDS RYS SAS SFS STS SUM BLR GOR NHS SHR QLL RVR SVR
SAB BKS BLG GSF LMB LOS OSF ROB SMB BAD BLD JOD LOP SLD;
LINES;
STATION1 6 164 42 0 13 0 10 0 8 1 19 45 0 0 34 5 35 5 2 8 0 4 2 1 0 35 1 30 143 4 1 0 25 5
STATION2 7 90 33 0 58 0 11 22 89 22 24 32 0 0 0 2 0 2 0 1 3 10 0 42 3 94 0 3 59 0 9 6 0 0
STATION3 0 6 29 0 10 0 10 4 5 4 3 69 0 4 35 10 8 1 2 1 0 13 0 6 2 26 0 15 195 0 1 3 51 6
STATION4 4 6 3 1 14 0 0 0 15 1 2 14 0 0 36 7 2 4 0 3 0 9 2 3 0 39 8 31 151 1 0 0 42 7
STATION5 26 432 44 15 36 7 2 0 8 5 23 51 4 1 9 2 35 17 0 14 3 10 1 7 8 48 31 27 165 0 4 2 2 4 2
STATION6 5 194 35 0 13 0 0 2 35 1 8 14 1 0 55 7 22 14 5 4 0 6 0 1 0 37 3 62 204 1 0 0 7 2
;
PROC FASTCLUS DATA = SPECIES MAXCLUSTERS = 3 OUT = CLUSTER SHORT;
VAR LOG GZS BLM BUM CAP HOC MIS RDS RYS SAS SFS STS SUM BLR GOR NHS SHR QLL RVR SVR SAB BKS BLG
GSF LMB LOS OSF ROB SMB BAD BLD JOD LOP SLD;
ID STATION;
PROC SORT; BY CLUSTER;
PROC PRINT NOOBS; VAR CLUSTER STATION DISTANCE;
RUN;

```

Program Output

Table Output for SAS' FASTCLUS procedure for *k*-means clustering of fish abundance data for six stations on the Kankakee River, Illinois (Box 15.1). The number of clusters, $k = 3$, is specified by the MAXCLUSTERS command. Settings for other commands (Replace = FULL, Radius = 0, and Maxiter = 1) are left at SAS/STAT default settings; see SAS manual for details (SAS Institute 2004). Note that *k*-means clustering procedure should be used only when the number of assemblages (samples) exceeds 30 (illustrated here with six samples). The root mean square between members within-group standard deviation is given by RMS SD. Maximum distance from seed to observation is the greatest difference between a random-number seed to an observation in that cluster. Criterion based on final seeds = 14.8912. Overall R^2 is a measure of predictability of the fish assemblage within a cluster. Further explanation of table values follows.

Cluster Summary

Cluster	RMS frequency	SD	Maximum distance from seed to observation	Nearest cluster	Distance between cluster centroids
1	1		0	3	319.3
2	2	22.4843	92.7052	3	129.7
3	3	20.3094	125.7	2	129.7

Statistics for Variables

Variable	Total STD	Within STD	R^2	$R^2/(1 - R^2)$
Longnose gar (Other 32 species)	9.14330	2.89636	0.939793	15.609272
Slenderhead darter	2.60768	3.32499	0.024510	0.025126
Overall	31.24607	21.05932	0.727448	2.669025
Pseudo F -statistic	4.00 ^a			
Approximate expected overall R^2	0.7274 ^a			
Cubic clustering criterion ^b				

Cluster Membership

Cluster	Station	Distance
1	5	0.000
2	2	92.705
2	4	92.705
3	1	61.506
3	3	125.706
3	6	91.995

^a The two values are invalid for correlated variables.

^b None calculated because of small sample size.

The output summary indicates that three clusters were specified (MAXCLUSTERS = 3). The remaining variables in the summary are additional clustering options; see SAS manual for details. In the first part of the table, distance between cluster centroids is used to examine the overall relationship among cluster members. Clusters 2 and 3 have the smallest distances, and hence, their members are more similar than those in cluster 1. The statistics for variables portion of the table is used to examine the change in R^2 with the number of clusters (k). The overall R^2 (0.727) and cubic clustering criterion (none was calculated because of small sample size) would be used to estimate the optimal number of clusters for the classification (Figure 15.7). The final portion of the table contains cluster membership and members' Euclidean distances from the cluster centroid (i.e., the cluster mean), which can be used to examine similarity among cluster members. For example, cluster 3 contains stations 1, 3, and 6. Among these, stations 1 and 6 are the most similar because they have the smallest distances (61 and 92).

clusters), and hence, ordination is a better technique for examining relationships among assemblages. Unconstrained ordination is generally not used for classification but can be useful for verifying those relationships indicated by cluster analysis and for suggesting alternative classifications.

There are several unconstrained ordination methods, and each has its advantages and limitations. The most commonly used techniques are principal component analysis (PCA) and nonmetric multidimensional scaling (NMDS). Of the two, PCA produces the most detailed and quantifiable measures of the relationships among assemblages (Cliff 1987). However, in an extensive evaluation of ecological ordination methods, Minchin (1987) found that PCA performed very poorly when relationships were nonlinear, whereas NMDS was the most robust technique. Further, he suggested that NMDS was the most appropriate ordination method for ecological applications. Relationships among fish assemblages are likely to be nonlinear to varying degrees. Hence, the relationships indicated by PCA could be biased. To maximize ecological insight and minimize the potential for bias, we recommend that fisheries scientists use both techniques and compare their results.

15.4.3.1 *Principal Component Analysis*

Principal component analysis reduces species (relative) abundances into linear combinations (i.e., principal components) that are uncorrelated with each other (Stevens 1992). It emphasizes the variation among assemblages, rather than similarities, and assumes that approximately linear relationships exist among fish assemblages. Prior to PCA, species abundances are standardized by estimating the correlation or covariance between species. Fisheries scientists should always standardize using correlation, which is the default for most statistical software. Based on the correlation matrix, PCA begins by finding the linear combination of species-specific abundances that accounts for the greatest amount of variation among samples. This linear combination is called the first principal component, which is simply a linear regression using (standardized) species abundances as predictors of a principal component score. The collection of the corresponding linear regression coefficients (one coefficient per species) is called the eigenvector of the first principal component, and the amount of variance explained is estimated as the eigenvalue.

Following the estimation of the first principal component, PCA finds a second linear combination (regression) of species abundances that accounts for the largest amount of the remaining variance (i.e., after the variance attributable to the first component is removed) and is pairwise uncorrelated with the first component. This is the second principal component, which also has an eigenvector and eigenvalue. The third principal component then is derived to be uncorrelated to the first two components and accounts for the third largest amount of variance. The process of creating uncorrelated linear combinations (principal components) is continued with each component accounting for the remaining variation until none remains. Thus, PCA attempts to summarize the pattern of variation among assemblages with a smaller number of components (compared to the number of

species) that accounts for most of the variance in the original data set. For most applications, this can be accomplished with fewer than five principal components.

The relationships among assemblages are determined by examining plots of principal component scores, with similar assemblages being located close together and dissimilar ones farther apart. For example, fish assemblages in stations 1 and 6 on the Kankakee River are similar to one another but are very different from those at stations 2 and 5 (Figure 15.8). Principal component scores are computed

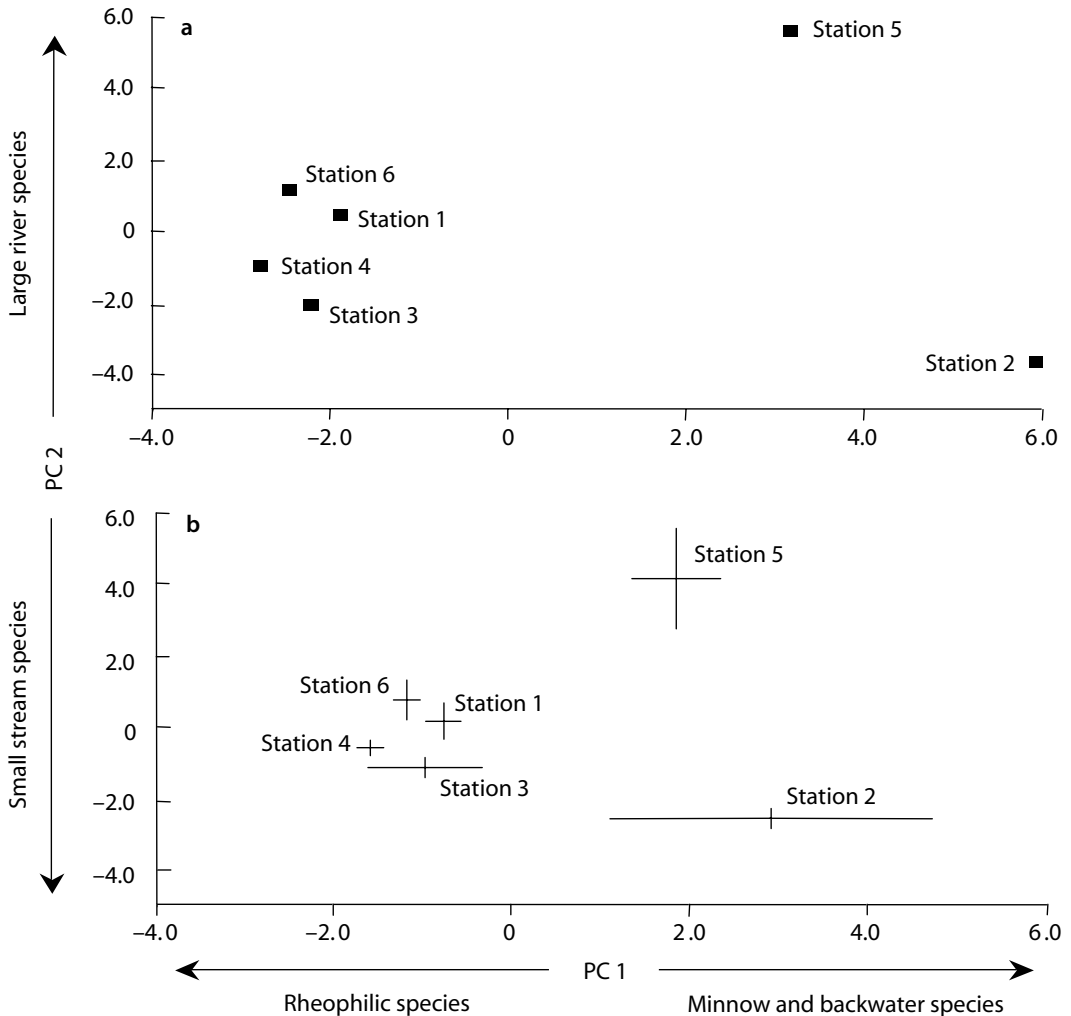


Figure 15.8 Plots of the first two principal components (PCs) of fish assemblages at six stations on Kankakee River, Illinois. The upper plot (a) was created using the PC scores from an analysis with summary data in Box 15.1. The bottom figure (b) was created using the PC scores from an analysis using eight sequential fish samples collected at each station (see Box 15.1 for summed data). The crossbars in panel (b) represent the mean and standard errors of sample scores for each station and can be used to examine the degree of overlap in assemblage structure. Principal component axes were interpreted using component loadings (Box 15.11).

for each assemblage (or sample) using the principal component eigenvectors (i.e., linear regression coefficients), and hence, each assemblage has a score for each principal component.

Each principal component is interpreted for ecological meaning by examining the principal component loadings, which are the Pearson's correlation coefficients between the component scores and species-specific abundances. These are also referred to as factor loadings or simply factors. Large loadings (in absolute value) are interpreted as having the greatest influence on that component. For example, if three species loaded high on a particular component, the component would be interpreted in terms of some characteristic these species have in common, such as taxonomy, habitat use, or environmental tolerances. Identifying the loadings of components to use for interpretation, however, is somewhat subjective. Because loadings are Pearson correlations, many fisheries scientists interpret only loadings that are statistically significant. These statistical tests are significantly influenced by sample sizes (Cliff 1987). For example, a species with a 0.20 loading would share only 4% of the variance with the principal component but could be statistically significant when sample sizes are large. Rather, a good general rule-of-thumb is to use species with loadings greater than $|0.4|$ for interpretation when there are small numbers of species in the analysis (<20) and $|0.6|$ for more species-rich assemblages (Stevens 1992). An example SAS program and output for PCA are provided in Box 15.11.

Theoretically, the maximum number of principal components in an analysis of fish assemblages is equal to the number of variables (species) analyzed. Thus, researchers must choose how many components to include in their analysis. There is no perfect criterion for selecting the number of components to retain. However, several techniques have been developed to assist the analyst. The most widely used of these is called the scree test (Cattell 1966). With this method, eigenvalues are plotted against their corresponding principal component number (i.e., first component = 1, second = 2, etc.). Eigenvalues generally decrease rapidly (steep slope) with increasing component number and then level off (shallow slope). The transition between these two slopes is known as the break point (Figure 15.9), and the components with eigenvalues above the breakpoint are retained for subsequent analysis. Another widely used approach is to retain all of the components with eigenvalues greater than 1 (Kaiser 1960). This method virtually ensures that important components will be retained but also tends to include additional components that have little explanatory value (Hakstian et al. 1982). A third criterion is based on practical considerations. There are two difficulties associated with using large numbers of principal components for examining the relationships among assemblages. First, three dimensions are difficult to display in a single figure and are substantially more difficult to comprehend than two. Second, four or more dimensions render ordination practically useless as a method of understanding complex relationships. Thus, we recommend that only two components be retained in fish assemblage analyses if these account for at least 70% of the variance; otherwise three components should be retained. However, if three components account for less than one half the variation, we suggest refraining from using PCA and consider NMDS as a more parsimonious alternative.

Box 15.11 Principal Component Analysis

The following SAS program performs principal component analysis (PCA) on the summary fish abundance data from six stations of the Kankakee River, Illinois (Box 15.1). Component loadings are estimated for the first three principal components, and scores are plotted for the first two components.

Program

```

OPTIONS PS = 60 LS=78;
DATA SPECIES;
INPUT STATION $ LOG GZS BLM BUM CAP HOC MIS RDS RYS SAS SFS STS SUM BLR GOR NHS SHR QLL RVR SVR
SAB BKS BLG GSF LMB LOS OSF ROB SMB BAD BLD JOD LOP SLD;
LINES;
STATION1 6 164 42 0 13 0 10 0 8 1 19 45 0 0 34 5 35 5 2 8 0 4 2 1 0 35 1 30 143 4 1 0 25 5
STATION2 7 90 33 0 58 0 11 22 89 22 24 32 0 0 0 2 0 2 0 1 3 10 0 42 3 94 0 3 59 0 9 6 0 0
STATION3 0 6 29 0 10 0 10 4 5 4 3 69 0 4 35 10 8 1 2 1 0 13 0 6 2 26 0 15 195 0 1 3 51 6
STATION4 4 6 3 1 14 0 0 0 15 1 2 14 0 0 36 7 2 4 0 3 0 9 2 3 0 39 8 31 151 1 0 0 42 7
STATION5 26 432 44 15 36 7 2 0 8 5 23 51 4 1 9 2 35 17 0 14 3 10 1 7 8 48 31 27 165 0 4 2 2 4 2
STATION6 5 194 35 0 13 0 0 2 35 1 8 14 1 0 55 7 22 14 5 4 0 6 0 1 0 37 3 62 204 1 0 0 7 2
;
PROC PRINCOMP DATA=SPECIES OUT = PRIN;
VAR LOG GZS BLM BUM CAP HOC MIS RDS RYS SAS SFS STS SUM BLR GOR NHS SHR QLL RVR SVR SAB BKS BLG
GSF LMB LOS OSF ROB SMB BAD BLD JOD LOP SLD;

PROC CORR NOPRINT OUTP = LOADING NOSIMPLE NOPROB;
VAR PRIN1 PRIN2 PRIN3;
WITH LOG GZS BLM BUM CAP HOC MIS RDS RYS SAS SFS STS SUM BLR GOR NHS SHR QLL RVR SVR SAB BKS
BLG GSF LMB LOS OSF ROB SMB BAD BLD JOD LOP SLD;
TITLE '*** COMPONENT LOADINGS ***';
PROC PRINT DATA = LOADING NOOBS; WHERE _NAME_ NE '';

PROC PLOT DATA=PRIN; PLOT PRIN2 * PRIN1 = '*' $ STATION /;
TITLE '*** PRINCIPAL COMPONENT SCORES ***';
RUN; TITLE; QUIT;

```

Program Output

Table Partial output of SAS program for principal component (PC) analysis. See “Component Loadings” at end of table for species abbreviations. Component loadings are Pearson correlation coefficients.

Simple Statistics (Partial)

	LOG	GZS	BLM	BUM	CAP
Mean	8.000000000	148.6666667	31.00000000	2.666666667	24.00000000
SD	9.143303561	159.1963149	14.81890684	6.055300708	19.17289754

Correlation Matrix (Partial)

	LOG	GZS	BLM	BUM	CAP	HOC	MIS	RDS	RYS
LOG	1.0000	0.9211	0.4812	0.9609	0.4438	0.9644	-0.3062	-0.1519	-0.1120
GZS	0.9211	1.0000	0.6948	0.8522	0.2727	0.8719	-0.3171	-0.2574	-0.1588
BLM	0.4812	0.6948	1.0000	0.3722	0.2583	0.4298	0.3477	0.0687	0.0333

(Box continues)

Box 15.11 (continued)**Eigenvalues of the Correlation Matrix**

PC	Eigenvalue	Difference	Proportion	Cumulative
1	13.4751939	3.2555850	0.3963	0.3963
2	10.2196089	5.5254938	0.3006	0.6969
3	4.6941151	1.7841269	0.1381	0.8350
4	2.9099883	0.2088945	0.0856	0.9206
5	2.7010937	2.7010937	0.0794	1.0000
6	0.0000000	0.0000000	0.0000	1.0000
7	0.0000000	0.0000000	0.0000	1.0000
(Remaining 27 PCs)				

Eigenvectors

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
LOG	0.148222	0.259093	0.030061	-0.029427	-0.064273	-0.019554	-0.064777
GZS	0.104941	0.279880	-0.065667	0.058686	0.087773	-0.009625	-0.043083
BLM	0.104793	0.139913	-0.073891	0.061684	0.477267	-0.031307	-0.012889
(Remaining species and 27 PCs)							

Component Loadings

SpeciesSpecies	Abbreviation	PC1	PC2	PC3
Longnose gar	LOG	0.54410	0.82827	0.06513
Gizzard shad	GZS	0.38522	0.89472	-0.14227
Bluntnose minnow	BLM	0.38468	0.44727	-0.16009
Bullhead minnow	BUM	0.40982	0.84812	0.30129
Common carp	CAP	0.98347	-0.11637	-0.08999
Hornyhead chub	HOC	0.42955	0.84878	0.29467
Mimic shiner	MIS	0.32226	-0.53181	0.20040
Redfin shiner	RDS	0.74136	-0.64720	-0.10607
Rosyface shiner	RYS	0.68056	-0.54750	-0.41512
Sand shiner	SAS	0.87490	-0.48009	-0.02615
Spotfin shiner	SFS	0.81830	0.28838	-0.28268
Striped shiner	STS	0.12277	0.11097	0.73543

15.4.3.2 *Multidimensional Scaling*

Nonmetric multidimensional scaling models the relationships among two or more assemblages in a specified number of dimensions based on their similarity or dissimilarity (Kruskal and Wish 1984). It is a robust ordination technique that does not require an assumption of normality or linearity of relationships among assemblages and can usually fit a model in fewer dimensions than can PCA. Nonmetric

Component Loadings (*continued*)

SpeciesSpecies	Abbreviation	PC1	PC2	PC3
Suckermouth minnow	SUM	0.35570	0.89406	0.15593
Black redhorse	BLR	-0.18744	-0.11634	0.87537
Golden redhorse	GOR	-0.92506	0.00075	-0.23981
Northern hog sucker	NHS	-0.84410	-0.32879	0.33545
Shorthead redhorse	SHR	-0.09627	0.83134	-0.22237
Quillback	QLL	0.08542	0.87966	-0.25527
River redhorse	RVR	-0.55565	0.01944	-0.41357
Silver redhorse	SVR	0.17510	0.94825	-0.02850
Smallmouth buffalo	SAB	0.96851	0.21751	0.08016
Brook silverside	BKS	0.28808	-0.23178	0.86445
Bluegill	BLG	-0.32942	0.30074	-0.14319
Green sunfish	GSF	0.84888	-0.52367	-0.05485
Largemouth bass	LMB	0.68457	0.57695	0.43733
Longear sunfish	LOS	0.89030	-0.34000	-0.26148
Orangespotted sunfish	OSF	0.31102	0.86433	0.25246
Rock bass	ROB	-0.59897	0.43907	-0.52590
Smallmouth bass	SMB	-0.74860	0.43763	0.23112
Banded darter	BAD	-0.44120	0.08443	-0.52494
Blackside darter	BLD	0.97080	-0.23332	-0.01678
Johnny darter	JOD	0.80895	-0.46298	0.30348
Logperch	LOP	-0.70992	-0.34408	0.55491
Slenderhead darter	SLD	-0.69835	0.16420	0.50953

The first two principal components accounted for 69.7% of the variation among fish assemblages and were retained for analysis. The first component accounted for 39.6% of the variation and loaded heavily and positively on minnows (RDS, RYS, SAS, SFS) and backwater species (CAP, SAB, GSF, LMB, LOS) and negatively on rheophilic species (GOR, NHS, SMB, LOP, SLD). The second component accounted for 30.1% of the variation and loaded heavily and positively on large-river species (LOG, GZS, BUM, HOC, SUM, SHR, QLL, SVR, OSF) and negatively on small stream-dwelling species (RDS). The plot of the first two principal components (Figure 15.8a) indicates that assemblages at stations 2 and 5 had larger numbers of minnow and backwater species, whereas stations 1, 3, 4 and 6 contained greater numbers of rheophilic species. Station 5 also included greater numbers of large-river species, and station 2 had greater numbers of small stream-dwelling species.

multidimensional scaling begins with a matrix of resemblance coefficients (section 15.4.1) and uses an iterative procedure and algorithm to find the set of coordinates for each assemblage that, when plotted, most closely approximates the relationships indicated by the resemblance matrix. For example, similar assemblages should be located closer together in an NMDS plot and dissimilar assemblages farther apart. The number of coordinates depends upon the specified number of dimensions (e.g., two dimensions = two coordinates). The degree of

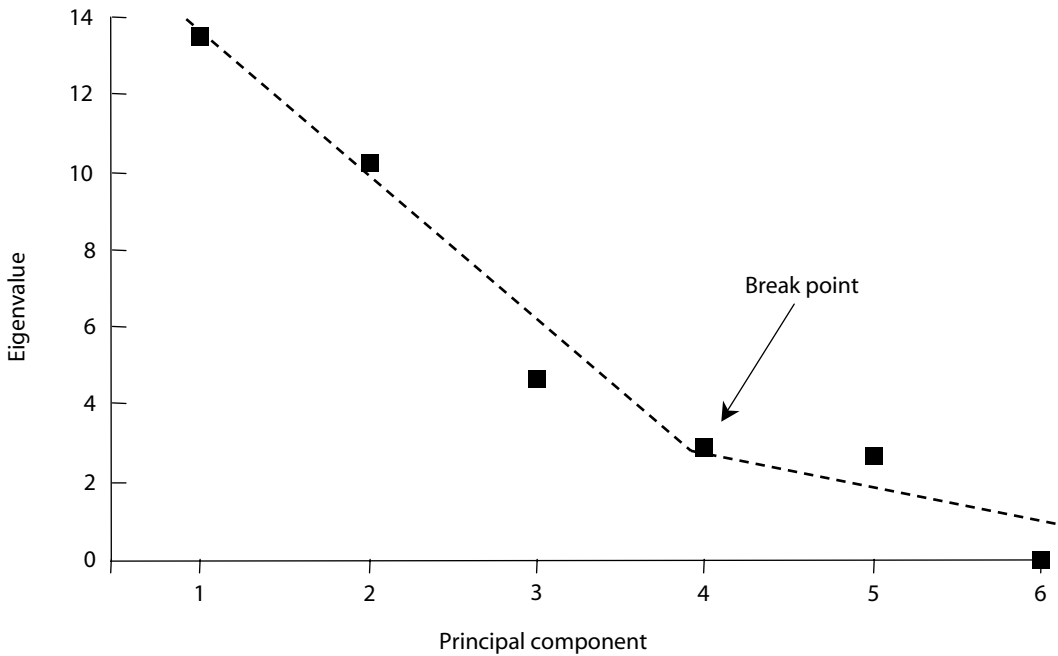


Figure 15.9 Scree plot of principal component eigenvalues from the analysis of Kankakee River fish assemblages (Boxes 15.1 and 15.11). Broken line is shown for illustration, and arrow indicates break point. Based on the scree test, principal components 1–3 would be retained for analysis.

correspondence between the resemblance coefficients and an NMDS plot is measured as stress, with lower stress values indicating better fit. Thus, NMDS iteratively finds the best-fitting coordinates by minimizing stress.

There are several algorithms that can be used to fit NMDS coordinates (Green and Rao 1972). The calculation and theory behind each are technical and beyond the scope of this chapter. However, there are two practical methods fisheries scientists can use to examine the fit of an NMDS model and determine the optimal number of dimensions. The first is the final estimate of stress or other related measure (e.g., badness-of-fit in SAS Institute 2004) that is estimated during the final iteration. Low values indicate better fit, and NMDS models with stress greater than 0.15 should be considered suspect (Kruskal and Wish 1984). The second method is inspection of a Shepard diagram, which is a scatterplot of the estimated similarity (or dissimilarity) among assemblages as shown in the NMDS plot versus the observed (actual) similarities (Shepard 1963). A Shepard diagram for a good-fitting model should resemble a smooth curve or a straight line (Box 15.12). Diagrams that resemble a stair-step or L-shaped function indicate a poor-fitting model, due to an incorrect specification of number of dimensions or the use of an incorrect algorithm. An example SAS program and output for NMDS are provided in Box 15.12.

Box 15.12 Nonmetric Multidimensional Scaling

The following SAS program performs nonmetric multidimensional scaling (NMDS) analysis on the summary fish abundance data from six stations of the Kankakee River, Illinois (Box 15.1). The resemblance measure is percent similarity, the scaling algorithm is monotonic (LEVEL = ORDINAL), and two dimensions are specified (DIM = 2). For more options, consult the SAS manual (SAS Institute 2004).

Program

```

OPTIONS PS = 60 LS=78;
DATA PSI;
INPUT STATION $ STATION1 STATION2 STATION3 STATION4 STATION5 STATION6;
LINES;
STATION1 100.0 50.8 62.6 56.7 73.0 81.4
STATION2 50.8 100.0 38.0 37.3 49.6 46.5
STATION3 62.6 38.0 100.0 77.5 44.5 59.2
STATION4 56.7 37.3 77.5 100.0 43.7 64.3
STATION5 73.0 49.6 44.5 43.7 100.0 68.1
STATION6 81.4 46.5 59.2 64.3 68.1 100.0
;
PROC MDS DATA = PSI DIM=2 SHAPE=SQUARE LEVEL=ORDINAL OUT=MDSOUT OUTRES=RESID NONORM;
ID STATION;
DATA MDSOUT; SET MDSOUT; WHERE _TYPE_ = 'CONFIG';
PROC PLOT DATA=MDSOUT; PLOT DIM2 * DIM1 = '*' $ STATION /;
PROC PLOT DATA=RESID; PLOT FITDATA * FITDIST = '*' /;
TITLE "SHEPARD DIAGRAM";
RUN;
TITLE;
QUIT;

```

Program Output

Table Nonmetric multidimensional scaling analysis on the summary fish abundance data from six stations of the Kankakee River, Illinois (Box 15.1). The resemblance measure is percent similarity, the scaling algorithm is monotonic (LEVEL = ORDINAL), and two dimensions are specified (DIM = 2). Other command settings are SAS defaults (Shape = SQUARE, Condition = MATRIX, Coef = IDENTITY, Formula = 1, Fit = 1, Mconverge = 0.01, Gconverge = 0.01, Maxiter = 100, Over = 2, and Ridge = 0.00010). The convergence criteria were satisfied.

Iteration	Badness of-fit type	Criterion	Change in criterion	Convergence measures	
				Monotone	Gradient
0	Initial	0.1655			
1	Monotone	0.0237	0.1418	0.1608	0.7467
2	Gau-New	0.0153	0.008313		
3	Monotone	0.0101	0.005246	0.0115	0.7041
4	Gau-New	0.0101	0.0000406		
5	Monotone	0.003382	0.006670	0.009464	0.5672
6	Gau-New	0.002792	0.000590		0.0221
7	Gau-New	0.002792	6.8241 x 10 ⁻⁷		0.000167

(Box continues)

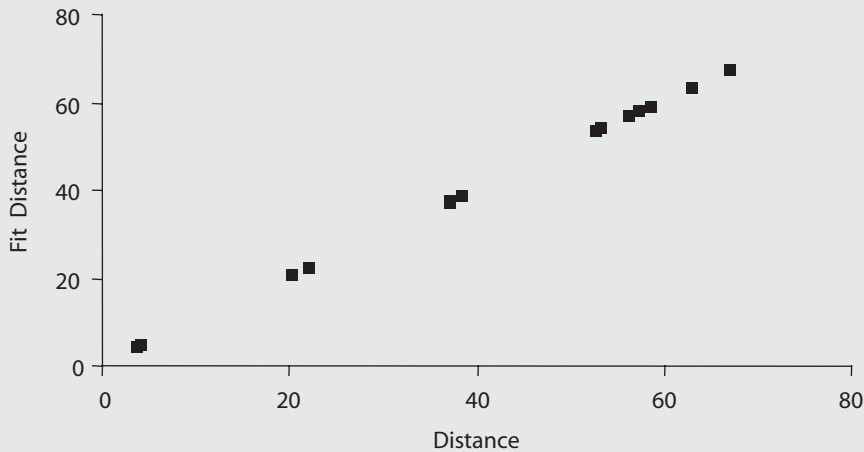
Box 15.12 (continued)


Figure Shepard diagram, which is a scatterplot of the estimated similarity (or dissimilarity) among assemblages as shown in the NMDS plot versus the observed (actual) similarities (Shepard 1963).

The badness-of-fit criteria (stress) declined smoothly among iterations and converged after seven iterations at 0.002792. The Shepard diagram displayed a smooth, linear relationship, which indicated a good fit for the two-dimensional NMDS model. The two-dimensional NMDS plot indicated that assemblages at stations 1 and 6 were most similar, as were those at stations 3 and 4 (Figure 15.10c). Fish assemblages at stations 2 and 5, however, differed from one another and from those at the other four stations. These relationships are also consistent with those indicated by PCA (Figure 15.8).

Similar to PCA, relationships among assemblages are determined by examining NMDS plots, and similar assemblages are located proximally and dissimilar assemblages distally. These relationships, however, are influenced by the characteristics of the resemblance measure, which is similar to hierarchical cluster analysis. For example, NMDS with Jaccard's similarity index suggests that fish assemblages at all stations on the Kankakee River are very different from one another, whereas NMDS with Kendall's tau suggests that assemblages at stations 1, 4, and 6 are most similar, and the percent similarity index suggests that the assemblage at station 4 is most similar to station 3 (Figure 15.10). Unlike PCA, NMDS has no formal or quantifiable means to interpret the various dimensions (i.e., no loadings). Thus, it requires insight and further examination of assemblage characteristics to determine the underlying pattern (gradient) associated with each dimension.

Nonmetric multidimensional scaling models can be fit with one to n dimensions, where n is the number of assemblages (samples). The optimal number of dimensions can be determined, similar to PCA, by examining a scree plot of stress

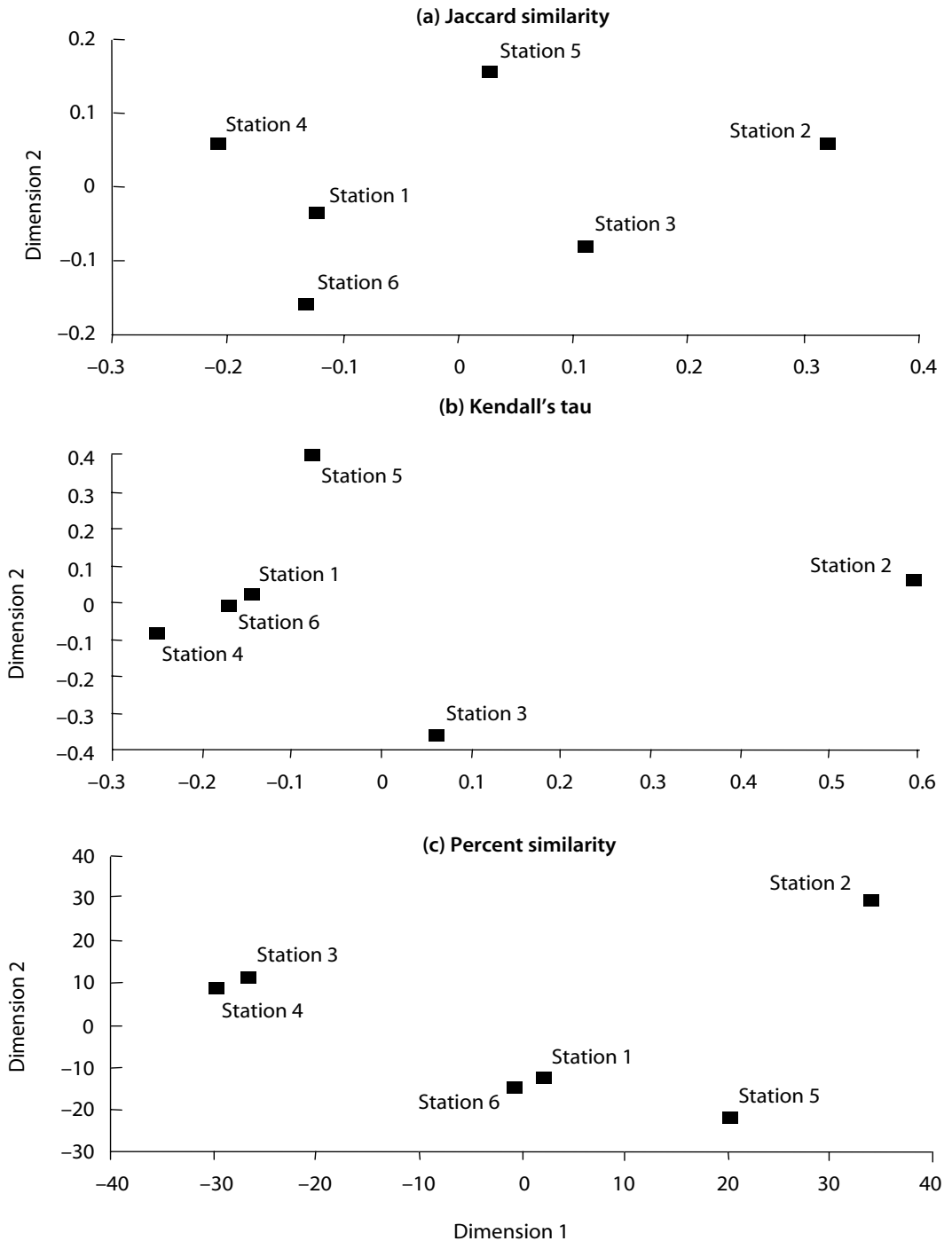


Figure 15.10 First two dimensions for nonmetric multidimensional scaling analysis of fish assemblages at six stations on the Kankakee River, Illinois (Box 15.1), with (a) Jaccard's similarity, (b) Kendall's tau, and (c) percent similarity resemblance matrices. Similar assemblages are located proximally and dissimilar assemblages are more distant.

against the corresponding number of dimensions or by examining a Shepard diagram (Green and Rao 1972). However, as with PCA, we suggest that fisheries scientists limit the maximum number of dimensions to three due to practical considerations, such as interpretability.

15.4.4 Other Multivariate Techniques—Discriminant Analysis

Fisheries scientists often need to determine how two or more fish assemblages differ and if assemblage composition is predictable. In these instances, assemblages are treated as discrete response categories (e.g., assemblages A, B, and C) with species (relative) abundances as their characteristics (i.e., predictors). Categorical data analysis is the generic term for a variety of statistical techniques for analyzing data with categorical responses (Agresti 1990). It can be used to find the species (or combination thereof) that best characterize an assemblage and can also be used to examine predictability of assemblage structure (Peterson and Rabeni 2001). In contrast to the other techniques in this chapter, categorical data analysis can be employed only when multiple samples (replicates) are collected from each assemblage. Hence, the quality of categorical data analysis is significantly influenced by sample size. To ensure reliable results, fisheries scientists should use categorical data analysis only when each assemblage has at least 20 samples (Cliff 1987).

There are several categorical data analysis techniques, the most widely used of which is discriminant analysis. Discriminant analysis is a linear statistical technique, requiring assumptions of normality and constant variance. It is relatively robust to minor violations of these assumptions (Stevens 1992) and should be appropriate for most practical applications in fisheries. However, biologists should consider alternative techniques when data are severely nonnormal and variances are heterogeneous. Fisheries scientists interested in alternative techniques may consult Agresti (1990) for logistic methods, Hand (1982) for nonparametric methods, and Breiman et al. (1984) for tree-based methods.

Discriminant analysis is a multivariate statistical technique, related to PCA, that reduces species (relative) abundances into linear combinations (i.e., discriminant functions) that are pairwise uncorrelated (Lachenbruch 1975; Klecka 1980). Discriminant analysis begins by finding the linear combination of species-specific abundances that accounts for the greatest differences among assemblages; hence, it is the best discriminator for separating (characterizing) the assemblages. This is in contrast to PCA, which simply accounts for the greatest amount of variation in the data. The linear combination is called the first discriminant function, which is a linear regression with species abundances as predictors. The amount of variance among assemblages that is explained by the discriminant function is estimated as the eigenvalue.

Similar to PCA, discriminant analysis then finds a second discriminant function that accounts for the largest amount of the remaining differences among assemblages and is pairwise uncorrelated with the first function. This is the second discriminant function, which is the second best discriminator for separating

the assemblages. The process of fitting discriminant functions then continues with consecutive discriminant functions representing smaller and smaller differences among assemblages. The maximum number of discriminant functions is determined by number of assemblages and the number of species analyzed. If the number of assemblages (k) is fewer than the number of species, the maximum number of discriminant functions is $k - 1$; otherwise, it is equal to the number of species. The statistical significance of each discriminant function can be determined with a residual testing procedure (Stevens 1992). Statistically significant functions are generally retained for interpretation.

Discriminant functions can be interpreted by two methods. The first is to examine the standardized discriminant function coefficients, which are estimated by multiplying each raw coefficient by the standard deviation of the corresponding species abundance. The second is to examine the discriminant function–variable correlations, which are analogous to PCA loadings. For both methods, the species with the larger coefficients and correlations (absolute value) are considered to have the greatest influence on the function, but they occasionally provide conflicting results. For example, a species can have a high standardized coefficient and a low correlation (and vice versa) for the same function. This generally occurs when some species abundances are strongly correlated. Discriminant function coefficients are partial regression coefficients. That is, they are estimated after the effects of the other species have been removed. Hence, they tend to be influenced by intercorrelations among species. The discriminant function–variable correlation, however, is a more direct estimate of the relationship between a species and the function and is generally more stable when sample sizes are smaller (Stevens 1992). Therefore, we recommend use of discriminant function–variable correlations to interpret discriminant functions and use of the standardized coefficients to determine which variables are redundant (i.e., correlated).

Fish assemblage characteristics are interpreted by examining plots of discriminant scores. These scores are computed for each replicate sample by means of the discriminant function coefficients and are usually averaged for each assemblage. Discriminant function scores are generally plotted in two dimensions. The separation among assemblages along a discriminant function axis corresponds to the degree to which they differ on a particular function.

Assemblage predictability. Discriminant analysis can also be used to classify samples into one or more groups (e.g., assemblages) based upon species composition and abundances, and it can be used, via a V -fold cross-validation procedure, to assess the accuracy of assemblage structure classifications (Peterson and Rabeni 2001). In this procedure, samples are randomly placed into V groups, the samples from one group are excluded, and a model is fit with the data in the remaining $V - 1$ groups. The excluded group's samples are then classified using the discriminant model. This procedure is repeated for each group, and the proportion of misclassifications, among groups, is used to assess the predictability of the assemblage structure. A special case of cross-validation occurs when V equals the total sample size, which is called "leave-one-out" cross-validation (Lachenbruch 1975). Although cross-validation is a useful technique for examining the accuracy of as-

semblage structure classifications, it is noteworthy that high classification errors can also result from poorly fitting models due to factors such as nonnormal data. Hence, fisheries scientists should consider examining the error rate for various models to ensure that misclassification errors are the result of unpredictable fish assemblage structure rather than a poor-fitting model.

Discriminant analysis example. We demonstrate discriminant analysis by use of our example data set from the Kankakee River, Illinois, where six sites were each sampled eight times (see Box 15.1 for summed data). In total, 3,995 individuals and 34 species were collected during the survey. The 48 samples (6 sites \times 8 samples) were used to assess differences among the assemblages at each station and to determine if assemblage structure was predictable. In Box 15.13, we present the SAS program used to perform the discriminant analysis of the fish assemblages and the associated output.

Discriminant analysis of those fish assemblages indicated that the first three of five functions were statistically significant and accounted for 93.3% of variance among assemblages. The minimal amount of variance explained by the remaining two functions suggested that they were redundant; consequently, they were dropped from the analysis. The first function discriminated among assemblages based on the abundance of longnose gar, bullhead minnow, largemouth bass, blackside darter, and golden redhorse and accounted for 53.2% of the variance. The second function accounted for 32.3% of the variation and discriminated among assemblages based on the abundance of logperch. The third function discriminated among assemblages based on the abundance of redbfin shiner, gizzard shad, shorthead redhorse, silver redhorse, and smallmouth bass and accounted for 7.9% of the variation among assemblages.

Biplots indicated that assemblages at stations 2 and 5 differed from the others with higher densities of longnose gar, bullhead minnow, largemouth bass, and blackside darter and lower densities of golden redhorse (Figure 15.11). Assemblages at stations 3 and 4 also tended to have higher densities of logperch than did those at stations 1 and 6, whereas the assemblage at station 2 could be differentiated from that of station 5 by having higher densities of redbfin shiner and lower densities of gizzard shad, shorthead redhorse, silver redhorse, and smallmouth bass. As expected, the relationships among assemblages indicated by the discriminant function biplots were virtually identical to those suggested by the principal component plots (Figure 15.8). This finding reflects the similarity in procedures used to calculate discriminant functions and principal components.

The leave-one-out cross-validation procedure indicated a poor overall classification error rate of 50%, which was lower than would be expected by random (83.3%). The greatest assemblage predictability was for stations 2 and 5, with classification error rates of 25%. The high classification error rates for assemblages at stations 1 and 4 (75%) suggested that their assemblages were relatively unpredictable. The fish assemblage samples from stations 3 and 4 were most often misclassified as one another, which suggested that they were the most similar assemblages.

Box 15.13 Discriminant Analysis

Program

```

OPTIONS PS = 60 LS=78;
DATA SPECIES;
INPUT STATION $ LOG GZS BLM BUM CAP HOC MIS RDS RYS SAS SFS STS SUM BLR GOR NHS SHR QLL RVR
SVR SAB BKS BLG GSF LMB LOS OSF ROB SMB BAD BLD JOD LOP SLD;
LINES;
(input data lines)
;
PROC DISCRIM DATA= SAMPLES NOCLASSIFY SHORT CANONICAL OUT = SCORE OUTSTAT= STATS;
CLASS STATION;
DATA STDCOEF; SET STATS; WHERE _TYPE_ = 'SCORE';
PROC TRANSPOSE DATA = STDCOEF OUT = STDCOEF; ID _NAME_;
PROC PRINT NOOBS;
TITLE '** STANDARDIZED CANONICAL DISCRIM FUNCTION COEFFICIENTS **';
DATA CORR; SET STATS; WHERE _TYPE_ = 'STRUCTUR';
PROC TRANSPOSE DATA = CORR OUT = CORR; ID _NAME_;
PROC PRINT NOOBS;
TITLE '** VARIABLE CANONICAL DISCRIM FUNCTION CORRELATIONS **';
PROC MEANS DATA = SCORE NOPRINT; BY STATION; VAR CAN1 CAN2 CAN3 CAN4 CAN5;
OUTPUT OUT = CANMEANS MEAN = CAN1 CAN2 CAN3 CAN4 CAN5;
PROC PRINT NOOBS;
TITLE '** MEAN CANONICAL SCORES **';
RUN;
TITLE;
QUIT;

```

Program Output

Table Eigenvalues for canonical discriminant functions based on fish abundance data from six stations of the Kankakee River, Illinois (Box 15.1). Eigenvalues estimate the amount of variance among assemblages that is explained by the discriminant function.

Function	Eigenvalue	Difference	Proportion	Cumulative
1	71.9663	28.2795	0.5315	0.5315
2	43.6868	33.0355	0.3227	0.8542
3	10.6513	5.3225	0.0787	0.9328
4	5.3288	1.5621	0.0394	0.9722
5	3.7667	0.0278	1.0000	

Table Residual test of discriminant functions testing the null hypothesis that the canonical discriminant functions in the current row and all that follow are 0.

Function	Likelihood ratio	Approximate F-value	Numerator df	Denominator df	P > F
1	0.00000087	4.60	170	49.838	<0.0001
2	0.00006367	3.32	132	42.446	<0.0001
3	0.00284502	2.14	96	33.825	0.0068
4	0.03314817	1.74	62	24	0.0672
5	0.20978877	1.63	30	13	0.1756

(Box continues)

Box 15.13 (continued)

Table Standardized canonical discriminant function coefficients (Can1–5), which are estimated by multiplying each raw coefficient by the standard deviation of the corresponding species abundance. Fish species abbreviations are given in Box 15.11.

Species	Can1	Can2	Can3	Can4	Can5
BAD	2.22457	-0.48972	0.64003	-0.79062	0.08280
BLG	-1.68944	-0.30942	-1.21861	1.09008	-0.81268
BLD	-0.98030	-0.68695	0.07823	0.57650	-0.92412
BLR	-1.17705	-0.17038	-0.79393	-0.72426	0.29653
BLM	-3.33386	-2.79803	-0.39095	-1.58555	0.49815
BKS	-1.49200	-1.21403	-0.76243	-0.05284	0.78281
BUM	-5.06659	2.34851	0.23292	1.26939	0.22113
CAP	-0.19464	-0.26258	0.75233	-0.29933	0.40622
GOR	2.07015	-1.73841	-0.17166	1.57722	-0.99228
GSF	-0.03456	-1.95688	-1.17854	-0.52994	-1.59839
GZS	3.41728	-2.95221	1.35361	-1.60174	-1.54082
HOC	1.53056	-1.61176	2.94328	-0.66275	0.21363
NHS	-4.99973	2.02565	-0.84181	-0.05209	1.14016
JOD	3.16295	-0.00467	0.71130	-0.66375	-0.43669
LOG	-0.85862	-1.16071	0.60182	-0.65685	0.30545
LMB	-4.06244	4.20083	-2.07917	0.50675	1.50131
LOP	1.89960	5.17398	0.23476	-0.36722	-1.63614
LOS	2.59106	-0.68042	-1.03481	1.71630	-0.57696
MIS	1.08908	1.69851	0.55843	1.01839	-2.34229
SHR	2.20270	-0.64959	0.60190	-0.76786	0.69662
OSF	-1.05600	0.03937	-0.54739	2.12078	-1.37153
QLL	-2.04450	0.26525	0.15850	1.02008	0.58319
RDS	1.05416	2.87170	-0.73884	-0.99056	1.55117
RVR	-2.45861	1.14824	-0.26552	0.00837	0.90398
ROB	-0.46791	-1.95681	-0.83695	-0.67286	0.56299
RYS	1.55350	-3.03223	-0.76696	2.59213	-1.64421
SAS	-4.88904	1.58556	0.67048	-1.08944	2.55940
SVR	-1.30226	0.15896	0.09574	-0.37498	-0.22452
SLD	0.45848	-2.07767	0.53864	-0.47588	-0.45735
SMB	4.38130	2.71061	2.16573	-0.02066	1.69134
SAB	-1.18708	-1.36536	-1.25310	0.41939	-0.07079
SFS	-2.65351	2.37492	0.55013	-0.41617	-0.50139
STS	1.88926	1.12735	0.66837	0.01674	0.07120
SUM	2.62921	0.93052	0.48263	0.23275	0.03670

Table Variable–canonical discriminant function correlations (Can1–5), or the discriminant function–variable correlations, which are analogous to PCA loadings. Fish species abbreviations are given in Box 15.11.

Species	Can1	Can2	Can3	Can4	Can5
BAD	0.23811	-0.14468	0.13118	-0.05207	-0.21766
BLG	0.11125	0.01246	0.14516	0.16810	-0.31712
BLD	-0.44303	-0.04790	-0.20957	-0.18865	-0.13599
BLR	0.00311	0.34322	0.03631	-0.14784	0.18479
BLM	-0.12122	-0.21289	0.30191	-0.33172	-0.02541
BKS	-0.13964	0.05261	-0.08348	0.04821	0.07028
BUM	-0.55716	0.14021	0.50691	0.21321	-0.00930
CAP	-0.32469	-0.00431	0.11495	-0.00423	-0.03166
GOR	0.61771	0.00697	0.26700	0.15737	0.17672
GSF	-0.39040	0.03887	-0.31092	-0.12480	-0.06229
GZS	-0.31260	-0.33744	0.45179	-0.11546	0.02747
HOC	-0.40317	0.08760	0.36608	0.09808	0.01048
NHS	0.28916	0.18871	-0.04219	0.04614	0.13469
JOD	-0.30252	0.14858	-0.26393	-0.22838	0.03947
LOG	-0.46128	-0.10398	0.40494	0.13456	-0.11298
LMB	-0.46566	0.16627	0.15313	-0.04932	0.04056
LOP	0.34406	0.54093	0.27623	0.12823	-0.09737
LOS	-0.10547	-0.04831	0.03102	0.09769	-0.13537
MIS	-0.06203	0.06283	-0.14418	-0.38620	-0.20545
SHR	0.14154	-0.16356	0.54160	-0.20291	0.00342
OSF	-0.42115	0.07686	0.46911	0.37731	-0.00603
QLL	-0.10668	-0.30767	0.40928	0.16711	0.24778
RDS	-0.22552	-0.07220	-0.46580	-0.24083	0.02768
RVR	0.24577	-0.12735	0.09575	-0.15821	0.24201
ROB	0.17531	-0.16398	0.45425	0.20168	0.20621
RYS	-0.11243	-0.30703	-0.25619	-0.09533	0.05399
SAS	-0.38583	-0.01101	-0.28355	-0.16861	-0.05566
SVR	-0.18831	-0.07547	0.52481	0.06790	-0.16480
SLD	0.36209	0.27481	0.06532	0.12263	-0.09039
SMB	0.34025	0.18790	0.41150	0.18409	0.23553
SAB	-0.40926	-0.03277	-0.03778	-0.04076	-0.06411
SFS	-0.33468	-0.20200	0.20987	-0.19623	-0.14918
STS	-0.07438	0.10649	0.13131	-0.18815	-0.16235
SUM	-0.34175	-0.01094	0.35967	0.10455	0.12402

(Box continues)

Box 15.13 (continued)**Table** Mean canonical scores (Can1–5). Canonical scores are computed for each replicate sample by means of the discriminant function coefficients and are averaged for each station assemblage.

Station	Can1	Can2	Can3	Can4	Can5
1	7.78886	-3.07963	2.56286	-2.12270	-2.64169
2	-8.98018	-3.55705	-5.12470	-1.34585	-0.77444
3	4.46916	10.80398	-0.87505	-1.87546	1.61638
4	5.14086	2.54201	-1.64292	4.16695	-1.14496
5	-13.00349	2.19170	4.37312	0.79536	0.06922
6	4.58479	-8.90100	0.70670	0.38170	2.87549

Table Cross-validation summary to assess the accuracy of assemblage structure classification.

Station and summary	Station						Total
	1	2	3	4	5	6	
Number of Observations and Percent Classified into Station							
1	2 25.00	0 0.00	1 12.50	4 50.00	0 0.00	1 12.50	8 100.00
2	0 0.00	6 75.00	1 12.50	0 0.00	0 0.00	1 12.50	8 100.00
3	0 0.00	1 12.50	3 37.50	4 50.00	0 0.00	0 0.00	8 100.00
4	3 37.50	0 0.00	2 25.00	2 25.00	0 0.00	1 12.50	8 100.00
5	0 0.00	2 25.00	0 0.00	0 0.00	6 75.00	0 0.00	8 100.00
6	2 25.00	0 0.00	0 0.00	1 12.50	0 0.00	5 62.50	8 100.00
Total observations	7	9	7	11	6	8	48
Error Count Estimates for Station							
Rate	0.7500	0.2500	0.6250	0.7500	0.2500	0.3750	0.5000

15.4.5 Graphical Techniques

There are a number of graphical techniques available with statistical computer applications to plot multivariate data, which are useful to describe and compare visually fish assemblage compositions. Such graphical techniques are helpful in examining broad trends among samples, detecting relationships among assemblages, and identifying outlier data. A convenient and simple graphical technique of this type is the scatterplot matrix. This graphical matrix is a series of two-dimensional biplots of the species abundances comparing two assemblages or samples,

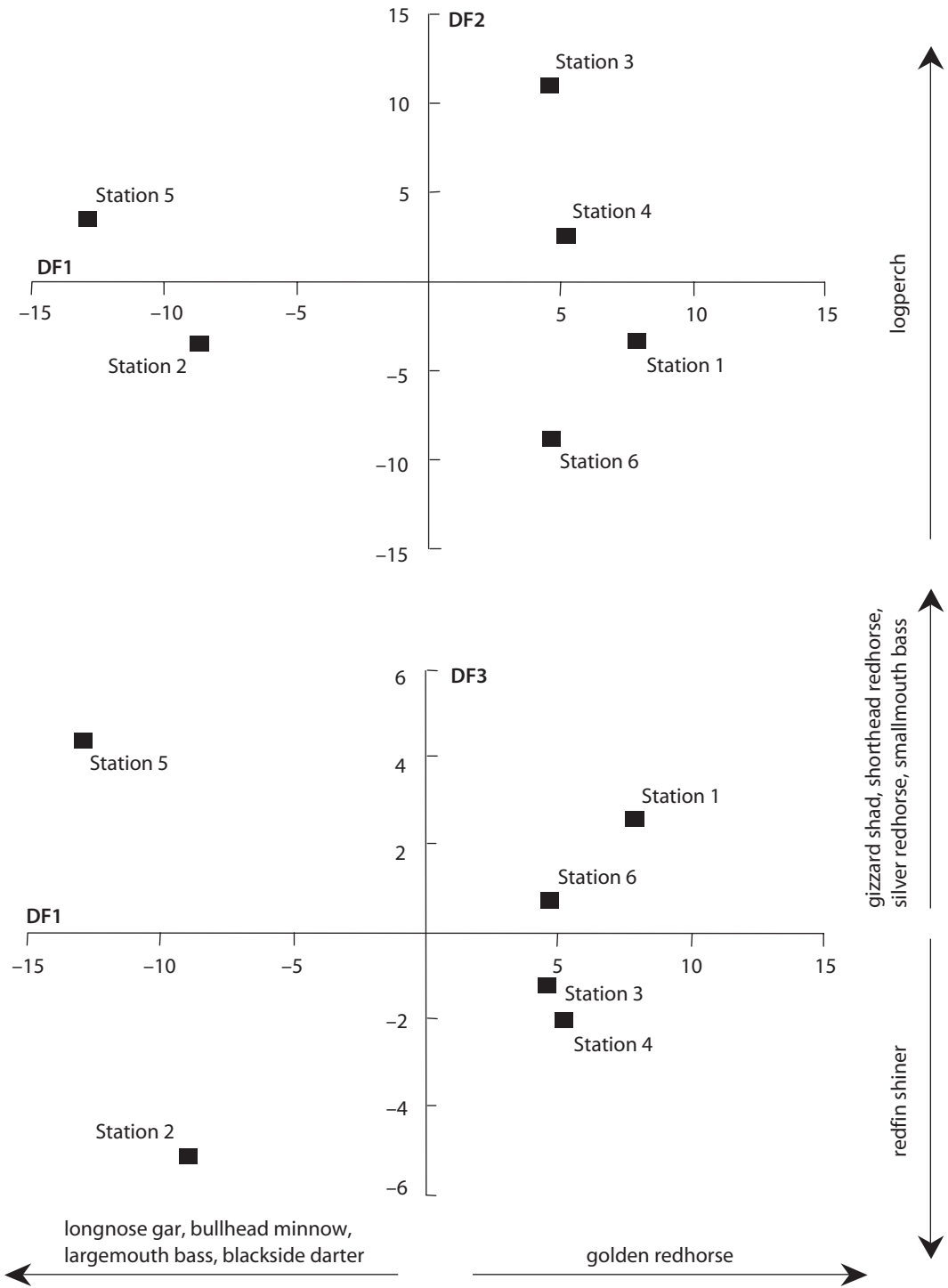


Figure 15.11 Discriminant function (DF) biplots for three functions of fish assemblages at six stations on the Kankakee River, Illinois (Boxes 15.1 and 15.13). Function interpretation and direction of positive influence (arrows) are shown at right and the bottom.

where each point on the plot represents the number of individuals (or density) of a species. When these biplots are fit with regression lines or ellipses and arranged in a matrix, trends in relationships among assemblages may be revealed and outlier species or assemblages can be identified.

Presented in Figure 15.12 are scatterplot matrices, including 95% bivariate normal density ellipses fit to each plot, that compare the six fish assemblages from the Kankakee River, Illinois example data (Box 15.1). Assemblage similarity is interpreted by examining the shape of the ellipse. Similar assemblages tend to have more linear, elongate ellipses, whereas a circular ellipse depicts dissimilar assemblages. An exploratory scatterplot matrix (Figure 15.12a) to identify similarities among assemblages can suggest a reordering of assemblage samples to more clearly reveal groups of similar assemblages (Figure 15.12b). This visual presentation suggests that assemblages at stations 1, 5, and 6 share a similar composition of fish abundances, as do the assemblages of stations 3 and 4, and that the station 2 assemblage is distinct from the others. This finding is virtually the same as that derived by multidimensional scaling, based on a percent similarity resemblance matrix (Figure 15.10c). Examination of individual points (one for each species) within the scatterplot matrix also indicates a number of outlier species that can explain differences among assemblages and may warrant additional analysis and interpretation.

Other graphical techniques that may be applicable to revealing attributes of, and relationships among, fish assemblages are Chernoff faces, star plots, sun-ray plots, and Andrews' plots (Johnson 1998). These techniques all share the properties that trends are relative, rather than absolute, and their detection depends upon the discerning eye and interpretive ability of the fisheries scientist; still, these techniques may reveal findings that could remain undisclosed by other more quantitative procedures.

■ 15.5 SUMMARY

Several common themes emerge from this chapter that characterize a general approach and provide guidance toward the description and comparison of fish assemblages. (1) Usually, more than one quantitative approach or technique is available to describe or compare fish assemblages. (2) The most appropriate approach depends on scientific objectives and data form, quality, and quantity. (3) Comparison of results from more than one approach may be useful to elucidate trends and overcome bias or artifacts of any single technique. (4) Analytical techniques should be selected prior to analyses, based on objectives and application, rather than posthoc conformity to expectations. (5) Most results related to fish assemblages are relative values that are meaningful in a comparative context rather than in an absolute sense. In such applications, absolute probabilities (P -values) and statistical significance (α -levels), to which many fisheries scientists are accustomed, are less applicable, and reliance upon them may confuse interpretation. This observation, however, does not excuse the scientist from a quantitative approach; on the contrary, intensive data description, exploration, and comparison

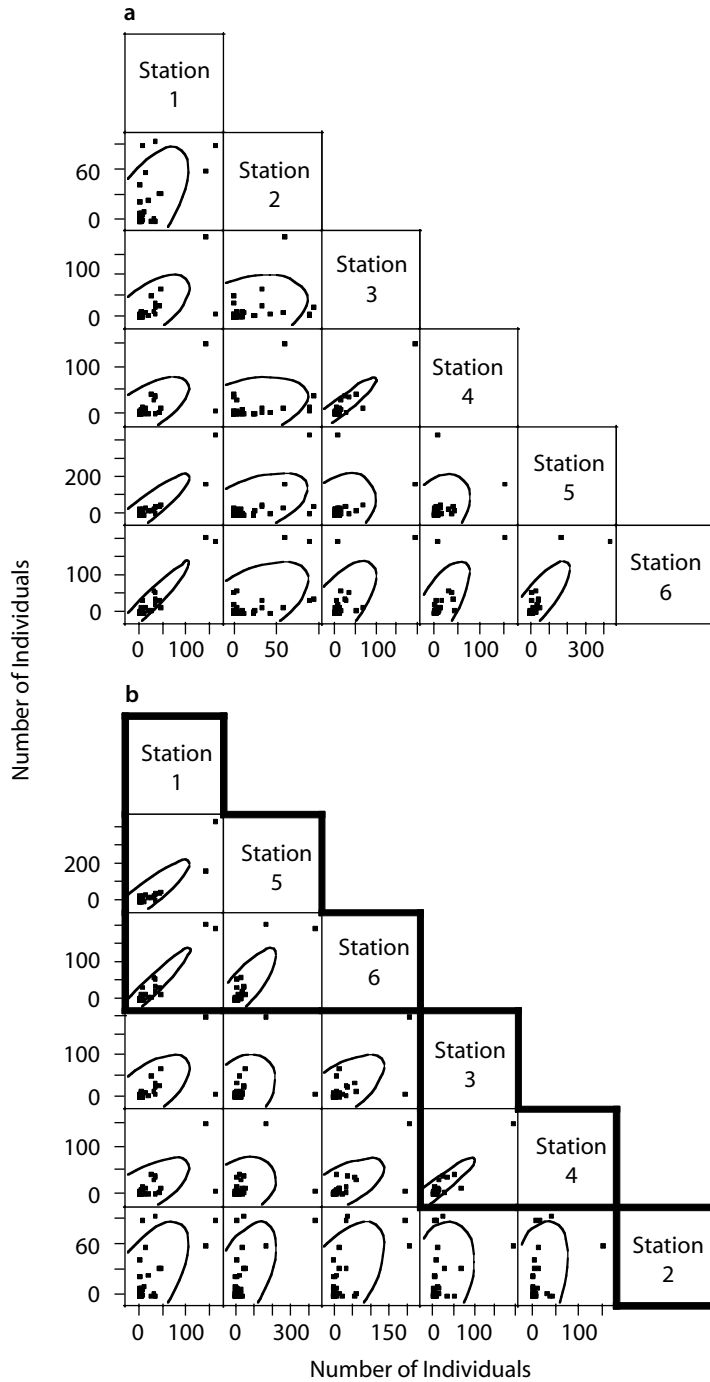


Figure 15.12 Scatterplot matrices of fish assemblages at six stations on the Kankakee River, Illinois (Box 15.1), with stations arranged by number (a) and according to similarity of assemblages (b). Curved-line enclosures represent the 95% bivariate normal density ellipse. Station plots enclosed by thick-lined boxes (b) depict similar fish assemblages.

are required to study fish at the community level. (6) Results of quantitative techniques are only valid if associated assumptions are not violated to a substantial degree. (7) The quality of results depends on quality of data. Discussion of data quality, logistic constraints, sampling bias and efficiency, analytical limitations, and other sampling and analytical concerns should not be avoided. Finally, the complexity of analyses at the community level precludes any strict protocol and allows for development of novel approaches that are limited only by the knowledge and creativity of the scientist; such quantitative methods and our understanding of them are likely to improve further in time.

■ 15.6 REFERENCES

- Agresti, A. 1990. *Categorical data analysis*. Wiley, New York.
- Angermeier, P. L., and I. J. Schlosser. 1989. Species–area relationships for stream fish. *Ecology* 70:1450–1462.
- Angermeier, P. L., and R. A. Smogor. 1995. Estimating number of species and relative abundances in stream-fish communities: effects of sampling effort and discontinuous spatial distribution. *Canadian Journal of Fisheries and Aquatic Sciences* 52:936–949.
- Angermeier, P. L., R. A. Smogor, and J. R. Stauffer. 2000. Regional frameworks and candidate metrics for assessing biotic integrity in mid-Atlantic highland streams. *Transactions of the American Fisheries Society* 129:962–981.
- Angermeier, P. L., and M. R. Winston. 1999. Characterizing fish community diversity across Virginia landscapes: prerequisite for conservation. *Ecological Applications* 9:335–349.
- Bailey, R. G. 1995. *Description of the ecoregions of the United States*, 2nd edition. U.S. Department of Agriculture, Forest Service, Miscellaneous Publication 1391, Washington, D.C.
- Bain, M. B., J. T. Finn, and H. E. Booke. 1988. Streamflow regulation and fish community structure. *Ecology* 69:382–392.
- Balon, E. K. 1975. Reproductive guilds of fishes: a proposal and definition. *Journal of the Fisheries Research Board of Canada* 32:821–864.
- Bayley, P. B., and D. C. Dowling. 1990. Gear efficiency calibrations for stream and river sampling. *Illinois Natural History Survey, Aquatic Ecology Technical Report 90/08*, Champaign.
- Bayley, P. B., and D. C. Dowling. 1993. The effects of habitat in biasing fish abundance and species richness estimates when using various sampling methods in streams. *Polskie Archiwum Hydrobiologii* 40:5–14.
- Bayley, P. B., and L. L. Osborne. 1993. Natural rehabilitation of stream fish populations in an Illinois catchment. *Freshwater Biology* 29:295–300.
- Bayley, P. B., and J. T. Peterson. 2001. Species presence for zero observations: an approach and an application to estimate probability of occurrence of fish species and species richness. *Transactions of the American Fisheries Society* 130:620–633.
- Benke, A. C. 1990. A perspective on America's vanishing streams. *Journal of the North American Benthological Society* 9:77–88.
- Berkman, H. E., and C. F. Rabeni. 1987. Effect of siltation on stream fishes. *Environmental Biology of Fishes* 18:285–294.

- Blackburn, T. M., J. H. Lawton, and J. N. Perry. 1992. A method for estimating the slope of upper bounds of plots of body size and abundance in natural animal assemblages. *Oikos* 65:107–112.
- Boulinier, T., J. D. Nichols, J. R. Sauer, J. E. Hines, and K. H. Pollock. 1998. Estimating species richness: the importance of heterogeneity in species detectability. *Ecology* 79:1018–1028.
- Bray, J. R., and J. T. Curtis. 1957. An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs* 27:326–349.
- Breiman, L., J. H. Friedman, R. Olshen, and C. J. Stone. 1984. *Classification and regression trees*. Wadsworth International, Belmont, California.
- Brown, M. L., and D. J. Austen. 1996. Data management and statistical techniques. Pages 17–62 in B. R. Murphy and D. W. Willis, editors. *Fisheries techniques*, 2nd edition. American Fisheries Society, Bethesda, Maryland.
- Cattell, R. B. 1966. The meaning and strategic use of factor analysis. Pages 174–243 in R. B. Cattell, editor. *Handbook of multivariate experimental psychology*. Rand McNally, Chicago.
- Chandler, J. R. 1970. A biological approach to water quality management. *Water Pollution Control* 69:415–421.
- Chao, A. 1984. Non-parametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics* 11:265–270.
- Cliff, N. 1987. *Analyzing multivariate data*. Harcourt Brace Jovanovich, New York.
- Colwell, R. K., and J. A. Coddington. 1994. Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London B* 345:101–118.
- Connor, E. F., and E. D. McCoy. 1979. The statistics and biology of the species area relationship. *American Naturalist* 113:791–833.
- Crowder, L. B. 1990. Community ecology. Pages 609–632 in C. B. Schreck and P. B. Moyle, editors. *Methods for fish biology*. American Fisheries Society, Bethesda, Maryland.
- Daniel, W. W. 1990. *Applied nonparametric statistics*, 2nd edition. PWS-Kent, Boston.
- Davis, W. S. 1995. Biological assessment and criteria: building on the past. Pages 15–29 in W. S. Davis and T. P. Simon, editors. *Biological assessment and criteria: tools for water resource planning and decision making*. Lewis Publishers, Boca Raton, Florida.
- Everitt, B. S. 1993. *Cluster analysis*, 3rd edition. Halsted Press, New York.
- Farris, J. S. 1969. On the cophenetic correlation coefficient. *Systematic Zoology* 18:279–285.
- Fausch, K. D., C. L. Hawkes, and M. G. Parsons. 1988. Models that predict standing crop of stream fish from habitat variables: 1950–85. U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station, General Technical Report PNW-GTR-213, Portland, Oregon.
- Fausch, K. D., J. R. Karr, and P. R. Yant. 1984. Regional application of an index of biotic integrity based on stream fish communities. *Transactions of the American Fisheries Society* 113:39–55.
- Fausch, K. D., J. Lyons, J. R. Karr, and P. L. Angermeier. 1990. Fish communities as indicators of environmental degradation. Pages 123–144 in S. M. Adams, editor. *Biological indicators of stress in fish*. American Fisheries Society, Symposium 8, Bethesda, Maryland.

- Field, J. G. 1970. The use of numerical methods to determine benthic distribution patterns from dredgings in False Bay. *Transactions of the Royal Society of South Africa* 39:183–200.
- Forbes, S. A. 1887. The lake as a microcosm. *Bulletin of the Peoria Scientific Association* 77–87. Reprinted, *Bulletin of the Illinois State Natural History Survey* 15(1925): 537–550.
- Gammon, J. R. 1976. The fish populations of the middle 340 km of the Wabash River. Purdue University, Water Resources Research Center Technical Report 86, Lafayette, Indiana.
- Gauch, H. G. 1982. *Multivariate analysis in community ecology*. Cambridge University Press, New York.
- Gerking, S. D. 1994. *Feeding ecology of fish*. Academic Press, San Diego, California.
- Glowacki, L., and T. Penczak. 2000. Impoundment impact on fish in the Warta River: species richness and sample size in the rarefaction method. *Journal of Fish Biology* 57:99–108.
- Gordon, N. D., T. A. McMahon, and B. L. Finlayson. 1992. *Stream hydrology: an introduction for ecologists*. Wiley, West Sussex, UK.
- Gorman, O. T., and J. R. Karr. 1978. Habitat structure and stream fish communities. *Ecology* 59:507–515.
- Gotelli, N. J., and G. R. Graves. 1996. *Null models in ecology*. Smithsonian Institution Press, Washington, D.C.
- Gray, J. S. 1987. Species-abundance patterns. Pages 53–67 *in* J. H. R. Gee and P. S. Giller, editors. *Organization of communities past and present*. Blackwell Scientific Publications, Oxford, UK.
- Green, P. E., and V. R. Rao. 1972. *Applied multidimensional scaling*. Holt, Rinehart, and Winson, New York.
- Green, R. H., and G. L. Vascotto. 1978. A method for the analysis of environmental factors controlling patterns of species composition in aquatic communities. *Water Research (Great Britain)* 12:583–590.
- Grossman, G. D., J. F. Dowd, and M. Crawford. 1990. Assemblage stability in stream fishes: a review. *Environmental Management* 14:661–671.
- Grossman, G. D., and M. C. Freeman. 1987. Microhabitat use in a stream fish assemblage. *Journal of Zoology* 212:151–176.
- Grossman, G. D., R. E. Ratajczak, M. Crawford, and M. C. Freeman. 1998. Assemblage organization in stream fishes: effects of environmental variation and interspecific interactions. *Ecological Monographs* 68:395–420.
- Hakstian, A. R., W. T. Rogers, and R. B. Cattell. 1982. The behavior of number-of-factors rules with simulated data. *Multivariate Behavioral Research* 17:193–219.
- Hall, C. A. 1972. Migration and metabolism in a temperate stream ecosystem. *Ecology* 53:586–604.
- Halliwell, D. B., R. W. Langdon, R. A. Daniels, J. P. Kurtenbach, and R. A. Jacobson. 1999. Classification of freshwater fish species of the northeastern United States for use in the development of indices of biological integrity, with regional applications. Pages 301–337 *in* T. P. Simon, editor. *Assessing the sustainability and biological integrity of water resources using fish communities*. CRC Press, Boca Raton, Florida.

- Hand, D. J. 1982. Kernel discriminant analysis. Research Studies Press, New York.
- Hankin, D. G., and G. H. Reeves. 1988. Estimating total fish abundance and total habitat area in small streams based on visual estimation methods. *Canadian Journal of Fisheries and Aquatic Sciences* 45:834–844.
- Hartigan, J. A. 1985. Statistical theory in clustering. *Journal of Classification* 2:63–76.
- Hayes, D. B., C. P. Ferreri, and W. W. Taylor. 1996. Active fish capture methods. Pages 193–220 in B. R. Murphy and D. W. Willis, editors. *Fisheries techniques*, 2nd edition. American Fisheries Society, Bethesda, Maryland.
- Heck, K. L., Jr., G. van Belle, and D. Simberloff. 1975. Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. *Ecology* 56:1459–1461.
- Horn, H. S. 1966. Measurement of overlap in comparative ecological studies. *American Naturalist* 100:419–429.
- Horton, R. E. 1945. Erosional development of streams and their drainage basins; hydrophysical approach to quantitative morphology. *Bulletin of the Geological Society of America* 56:275–370.
- Horwitz, R. J. 1978. Temporal variability patterns and the distributional patterns of stream fishes. *Ecological Monographs* 48:307–321.
- Hubert, W. A. 1996. Passive capture techniques. Pages 157–192 in B. R. Murphy and D. W. Willis, editors. *Fisheries techniques*, 2nd edition. American Fisheries Society, Bethesda, Maryland.
- Hughes, R. M. 1995. Defining acceptable biological status by comparing with reference conditions. Pages 31–47 in W. S. Davis and T. P. Simon, editors. *Biological assessment and criteria: tools for water resource planning and decision making*. Lewis Publishers, Boca Raton, Florida.
- Hughes, R. M., P. R. Kaufmann, A. T. Herlihy, T. M. Kincaid, L. Reynolds, and D. P. Larsen. 1998. A process for developing and evaluating indices of fish assemblage integrity. *Canadian Journal of Fisheries and Aquatic Sciences* 55:1618–1631.
- Hurlbert, S. H. 1971. The nonconcept of species diversity: a critique and alternative parameters. *Ecology* 52:577–586.
- Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* 54:187–211.
- Hutchinson, G. E. 1978. *An introduction to population ecology*. Yale University Press, New Haven, Connecticut.
- Hynes, H. B. N. 1970. *The ecology of running waters*. Liverpool University Press, Liverpool, UK.
- Johnson, D. E. 1998. *Applied multivariate methods for data analysts*. Duxbury Press, Pacific Grove, California.
- Johnson, D. H. 1999. The insignificance of statistical significance testing. *Journal of Wildlife Management* 63:763–772.
- Kaiser, H. F. 1960. The application of electronic computers to factor analysis. *Educational and Psychological Measurement* 20:141–151.
- Karr, J. R. 1981. Assessment of biotic integrity using fish communities. *Fisheries* (Bethesda) 6(6):21–27.
- Karr, J. R., and E. W. Chu. 1999. *Restoring life in running waters*. Island Press, Washington, D.C.

- Karr, J. R., K. D. Fausch, P. L. Angermeier, P. R. Yant, and I. J. Schlosser. 1986. Assessing biological integrity in running waters: a method and its rationale. Illinois Natural History Survey Special Publication 5, Champaign.
- Klecka, W. R. 1980. Discriminant analysis. Sage Publications, Beverly Hills, California.
- Kolkwitz, R., and M. Marsson. 1908. Ökologie der pflanzlichen Saprobien. Berichte der Deutschen Botanischen Gesellschaft 26a:505–519. (Translated 1967. Ecology of plant saprobia. Pages 47–52 in L. E. Keup, W. M. Ingram, and K. M. Mackenthum, editors. Biology of water pollution. U.S. Department of Interior, Federal Water Pollution Control Administration, Washington, D.C.)
- Krebs, C. J. 1998. Ecological methodology, 2nd edition. Benjamin/Cummings, Menlo Park, California.
- Krueger, C. C., and D. J. Decker. 1999. The process of fisheries management. Pages 31–59 in C. C. Kohler and W. A. Hubert, editors. Inland fisheries management in North America, 2nd edition. American Fisheries Society, Bethesda, Maryland.
- Kruskal, J. B., and M. Wish. 1984. Multidimensional scaling. Sage Publications, London.
- Kwak, T. J. 1993. The Kankakee River: a case study and management recommendations for a stream diverse in habitat, fauna, and human values. Pages 123–141 in L. W. Hesse, C. B. Stalnaker, N. G. Benson, and J. R. Zuboy, editors. Restoration planning for the rivers of the Mississippi River ecosystem. U.S. National Biological Survey Biological Report 19, Washington, D.C.
- Lachenbruch, P. A. 1975. Discriminant analysis. Collier Macmillan, New York.
- Leopold, L. B., M. G. Wolman, and J. P. Miller. 1964. Fluvial processes in geomorphology. Freeman, San Francisco.
- Lyons, J. 1992. The length of stream to sample with a towed electrofishing unit when fish species richness is estimated. North American Journal of Fisheries Management 12: 198–203.
- Matthews, W. J. 1998. Patterns in freshwater fish ecology. Chapman and Hall, New York.
- Miller, D. L., and 13 coauthors. 1988. Regional applications of an index of biotic integrity for use in water resource management. Fisheries (Bethesda) 13(5):12–20.
- Milligan, G. W., and M. C. Cooper. 1985. An examination of procedures for determining the number of clusters in a data set. Psychometrika 50:159–179.
- Minchin, P. R. 1987. An evaluation of the relative robustness of techniques for ecological ordination. Vegetatio 69:89–107.
- Morin, P. J. 1999. Community ecology. Blackwell Scientific Publications, Malden, Massachusetts.
- Morisita, M. 1959. Measuring of interspecific association and similarity between communities. Memoirs of the Faculty of Science of Kyushu University Series E 3:65–80.
- Mueller, D. W., and G. Sawitzki. 1991. Excess mass estimates and tests for multimodality. Journal of the American Statistical Association 86:738–746.
- National Research Council. 1992. Restoration of aquatic ecosystems: science, technology, and public policy. National Academy Press, Washington, D.C.
- Omernick, J. M. 1995. Ecoregions: a spatial framework for environmental management. Pages 49–62 in W. S. Davis and T. P. Simon, editors. Biological assessment and criteria: tools for water resource planning and decision making. Lewis Publishers, Boca Raton, Florida.

- Omernick, J. M., and R. G. Bailey. 1997. Distinguishing between watersheds and ecoregions. *Journal of the American Water Resources Association* 33:935–949.
- Osborne, L. L., R. W. Davies, and K. L. Linton. 1980. Use of hierarchical diversity indices in lotic community analysis. *Journal of Applied Ecology* 17:567–580.
- Paller, M. H. 1995. Relationships among number of fish species sampled, reach length surveyed, and sampling effort in South Carolina coastal plain streams. *North American Journal of Fisheries Management* 15:110–120.
- Palmer, M. W. 1990. The estimation of species richness by extrapolation. *Ecology* 71:1195–1198.
- Peet, R. K. 1974. The measurement of species diversity. *Annual Review of Ecology and Systematics* 5:285–307.
- Peterson, J. T. 1989. Kankakee River fishes of the Braidwood Station Aquatic Monitoring Area, August 1988. Illinois Natural History Survey, Aquatic Biology Technical Report 89/1, Champaign.
- Peterson, J. T., and C. F. Rabeni. 1995. Optimizing sampling effort for sampling warmwater stream fish communities. *North American Journal of Fisheries Management* 15: 528–541.
- Peterson, J. T., and C. F. Rabeni. 1996. Natural thermal refugia for temperate warmwater stream fishes. *North American Journal of Fisheries Management* 16:738–746.
- Peterson, J. T., and C. F. Rabeni. 2001. The relation of fish assemblages to channel units in an Ozark stream. *Transactions of the American Fisheries Society* 130:911–926.
- Pielou, E. C. 1975. *Ecological diversity*. Wiley, New York.
- Poff, N. L., and J. D. Allan. 1995. Functional organization of stream fish assemblages in relation to hydrological variability. *Ecology* 76:606–627.
- Quinn, J. W., and T. J. Kwak. 2003. Fish assemblage changes in an Ozark river after impoundment: a long-term perspective. *Transactions of the American Fisheries Society* 132:110–119.
- Rabeni, C. F., and R. B. Jacobson. 1999. Warmwater streams. Pages 505–528 in C. C. Kohler and W. A. Hubert, editors. *Inland fisheries management in North America*, 2nd edition. American Fisheries Society, Bethesda, Maryland.
- Rahel, F. J., J. D. Lyons, and P. A. Cochran. 1984. Stochastic or deterministic regulation of assemblage structure? It may depend on how the assemblage is defined. *American Naturalist* 124:583–589.
- Renkonen, O. 1938. Statistisch-ökologische Untersuchungen über die terrestrische Kaferwelt der finnischen Bruchmoore. *Annales Zoologici Societatis Zoologicae-Botanicæ Fennica Vanamo* 6:1–231.
- Rexstad, E. A., D. R. Miller, C. R. Flather, E. M. Anderson, J. W. Hupp, and D. R. Anderson. 1988. Questionable multivariate statistical inference in wildlife habitat and community studies. *Journal of Wildlife Management* 52:794–798.
- Ricker, W. E. 1975. Computation and interpretation of biological statistics of fish populations. *Fisheries Research Board of Canada Bulletin* 191.
- Rodgers, J. D., M. F. Solazzi, S. L. Johnson, and M. A. Buckman. 1992. Comparison of three techniques to estimate juvenile coho salmon populations in small streams. *North American Journal of Fisheries Management* 12:79–86.

- Romesburg, H. C. 1990. Cluster analysis for researchers. Krieger Publishing, Malabar, Florida.
- Root, R. B. 1967. The niche exploitation pattern of the blue-gray gnatcatcher. *Ecological Monographs* 37:317–350.
- Rosenburg, D. M., and V. H. Resh, editors. 1993. Freshwater biomonitoring and benthic macroinvertebrates. Chapman and Hall, New York.
- Sanders, H. L. 1968. Marine benthic diversity: a comparative study. *American Naturalist* 102:243–282.
- Sarle, W. S. 1983. Cubic clustering criterion. SAS Institute, SAS Technical Report A-108, Cary, North Carolina.
- SAS Institute. 2004. SAS/STAT 9 user's guide. SAS Institute, Cary, North Carolina. (Also available as SAS OnlineDoc 9.1.3.)
- Schlösser, I. J. 1982. Fish community structure and function along two habitat gradients in a headwater stream. *Ecological Monographs* 52:395–414.
- Shannon, C. E., and W. Weaver. 1949. The mathematical theory of communication. University of Illinois Press, Urbana.
- Shelford, V. E. 1929. Laboratory and field ecology: the responses of animals as indicators of correct working methods. Williams and Wilkins, Baltimore, Maryland.
- Shepard, R. N. 1963. Analysis of proximities as a study of information processing in man. *Human Factors* 5:33–48.
- Simberloff, D. S. 1972. Properties of the rarefaction diversity measurement. *American Naturalist* 106:414–418.
- Simon, T. P., editor. 1999a. Assessing the sustainability and biological integrity of water resources using fish communities. CRC Press, Boca Raton, Florida.
- Simon, T. P. 1999b. Introduction: biological integrity and use of ecological health concepts for application to water resource characterization. Pages 3–16 in T. P. Simon, editor. Assessing the sustainability and biological integrity of water resources using fish communities. CRC Press, Boca Raton, Florida.
- Simon, T. P. 1999c. Assessment of Balon's reproductive guilds with application to midwestern North American freshwater fishes. Pages 97–121 in T. P. Simon, editor. Assessing the sustainability and biological integrity of water resources using fish communities. CRC Press, Boca Raton, Florida.
- Simpson, E. H. 1949. Measurement of diversity. *Nature (London)* 163:688.
- Southwood, T. R. E., and P. A. Henderson. 2000. *Ecological methods*, 3rd edition. Blackwell Scientific Publications, Oxford, UK.
- Stevens, J. 1992. *Applied multivariate statistics for the social sciences*, 2nd edition. Lawrence Erlbaum Associates, Hillsdale, New Jersey.
- Strahler, A. N. 1957. Quantitative analysis of watershed geomorphology. *Transactions of the American Geophysical Union* 38:913–920.
- Suter, G. W., II. 1993. A critique of ecosystem health concepts and indexes. *Environmental Toxicology and Chemistry* 12:1533–1539.
- Ter Braak, C. J. F. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology* 67:1167–1179.
- Ter Braak, C. J. F., and I. C. Prentice. 1988. A theory of gradient analysis. *Advances in Ecological Research* 18:271–313.

-
- Vannote, R. L., G. W. Minshall, K. W. Cummins, J. R. Sedell, and C. E. Cushing. 1980. The river continuum concept. *Canadian Journal of Fisheries and Aquatic Sciences* 37: 130–137.
- Van Sickle, J. 1997. Using mean similarity dendrograms to evaluate classifications. *Journal of Agricultural, Biological, and Environmental Statistics* 2:370–388.
- Washington, H. G. 1984. Diversity, biotic and similarity indices: a review with special relevance to aquatic ecosystems. *Water Research (Great Britain)* 18:653–694.
- Wiley, M. J., L. L. Osborne, and R. W. Larimore. 1990. Longitudinal structure of an agricultural prairie river system and its relationship to current stream ecosystem theory. *Canadian Journal of Fisheries and Aquatic Sciences* 47:373–384.
- Winemiller, K. O., and K. A. Rose. 1992. Patterns in life history diversification in North American fishes: implications for population regulation. *Canadian Journal of Fisheries and Aquatic Sciences* 49:2196–2218.
- Wolda, H. 1981. Similarity indices, sample size, and diversity. *Oecologia* 50:296–302.
- Wright, S. J. 1988. Patterns of abundance and the form of the species-area relation. *American Naturalist* 131:401–411.

