American Fisheries Society Symposium 73:503–527, 2010

Cluster Analysis of Fish Community Data: "New" Tools for Determining Meaningful Groupings of Sites and Species Assemblages

Donald A. Jackson*, Steven C. Walker, and Mark S. Poos
Department of Ecology and Evolutionary Biology, University of Toronto
Toronto, Ontario M5S3G5, Canada

Abstract.—Community ecologists face the challenge of summarizing considerable amounts of information regarding species distributions and environmental conditions. Often, this challenge is met through the use of multivariate statistical approaches. Stream fish community ecologists, much like the broader ecological community, appear to favor the use of ordination methods over clustering approaches. One potential reason is due to the development of various tools to help us determine the interpretability or "significance" of ordination axes, whereas ecologists appear unfamiliar with the comparable tools available for examining cluster analysis. We use fish abundance data from two river systems to demonstrate several of these approaches. We demonstrate how the methods may be used to determine the relative strength of groups of sampling locations and species assemblages relative to the background variability. We contrast the methods to demonstrate their relative merits, both advantages and disadvantages, in studies commonly conducted by stream ecologists.

Introduction

Understanding factors regulating species abundance and distribution remains a challenge to ecologists. Population ecologists typically focus on one or two species to determine the mechanisms controlling population abundance and demography, yet often find it difficult to easily resolve the relationships in these systems. Community ecologists are faced with incorporating these challenges too, but they are multiplied many fold due to considering many species, often comparing these groups of species across numerous sites and a range of habitats. Working at the community level, we examine questions related to understanding the roles of biotic, abiotic, spatial, and historical factors in determining the structure of communities—a task requiring considerable amounts of data. We fortunately now have increasing abilities to obtain large amounts of environmental data using remote sensing and data loggers, but also face the challenges of having to summarize underlying relationships within such extensive data sets (Guisan and Zimmermann 2000). The simplest approaches to these problems involve summarizing sites with measures such as species richness, diversity, or measures of "biotic integrity." Yet such measures provide no indication of the types of species present, nor whether sites that have similar levels of richness, diversity or integrity even contain any species in common. Alter-

* Corresponding author: don.jackson@utoronto.ca

native approaches to characterizing the biota have included the use of functional traits (e.g., reproductive, feeding; see Frimpong and Angermeier 2010, this volume) or morphology (Tilman 1997). These latter approaches provide greater detail in contrasting the community from one site with those found at another site, but as we begin to consider many sites, we can rapidly become overwhelmed with the amount of data and simple tabular or visual comparisons become limited in their usefulness. Similarly, when examining environmental conditions related to these sampling locations, we are faced with numerous additional variables. These approaches are hence deceptively simple; they either fail to summarize much of the complexity of the data—as with species richness, diversity, or biotic integrity—or retain too much complexity that the underlying patterns remain unclear—as with tabular or visual comparisons. Developing a meaningful and complete summary of complex ecological data often requires more sophisticated multivariate statistical approaches.

Ecologists have used multivariate statistical methods, such as ordination and cluster analysis, as standard analytical tools for many decades. Such methods provide mechanisms to summarize large amounts of community and environmental data (e.g., Bowman et al. 2008; Winemiller et al. 2008; Poos et al. 2009). Community ecologists, including many fish ecologists, have been quick to adopt, develop, and evaluate many approaches to guide researchers in the interpretation of ordination solutions (e.g., Grossman et al. 1991; Jackson 1993; Peres-Neto et al. 2003, 2005) ranging from graphical methods to various statistical resampling methods. Some of these approaches have become standard tools for community ecologists. By contrast, there have been many methods developed to aid in the interpretation of cluster analyses in various fields (Milligan and Cooper 1985; Tonidandel and Overall 2004); yet community ecologists have virtually ignored such advances in their analyses. Ecologists still rely heavily on either simple visual assessments of the dendrograms from clustering or use approaches based on cut levels (i.e., identifying some arbitrary but specific level of resemblance that determines the point at which clusters will be defined and interpreted (Figure 1; e.g. Joergensen et al. 2005; Morris et al. 2006; Kwak and Patterson 2007; Mehner et al. 2007). Although there is a clear necessity to include visual assessments in the interpretation of cluster analyses, cluster solutions are known to be influenced by the choice of hierarchical clustering method; therefore, providing some more quantitative guidelines, to determine whether clusters are meaningful or methodological artifacts should improve our use of cluster analyses.

It is not clear why stream fish ecologists, or community ecologists in general, do not use these methods of assessing dendrogram structure, even though clustering approaches are commonly used. Using Web of Knowledge, we conducted a literature search, including papers published during 2007 and the first 8 months of 2008 using the keywords "fish community" and either "stream," "lake" or "marine." Using the papers identified, we examined each publication to determine whether multivariate analyses (ordination and cluster analysis) were used, and we categorized them based on whether they were stream, lake, or marine in the type of system studied. From a total of 123 papers that met the criteria, we found 29 papers summarizing stream fish communities, and 5 of these papers (17.2%) used clustering as either the only multivariate approach or in conjunction with ordination methods, whereas 11 (37.9%) papers used ordination approaches. The remaining 13 (44.8%) used neither approach. Community ecologists (n

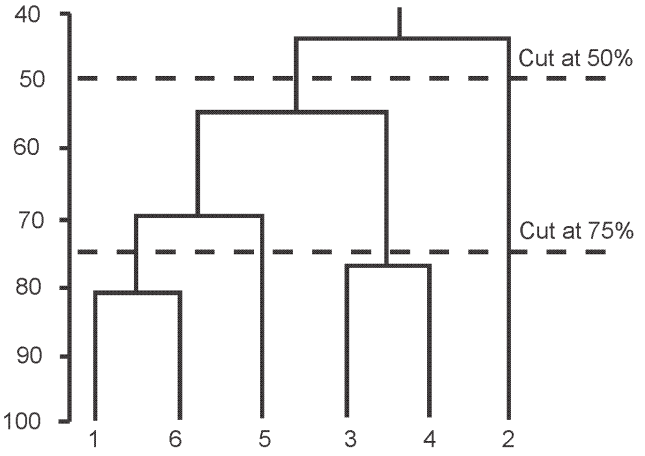

Figure 1. Example of cut-off classification redrawn from Kwak and Patterson (2007). A cut level is arbitrarily defined at which point all members that are grouped at a stronger resemblance level are considered to be a cluster. For example, stations 1 and 6 and stations 3 and 4 would represent two clusters at a cut level of 75% resemblance.

= 49) studying lake fishes used these methods only about half as frequently (6.1% and 18.4%, respectively), whereas marine ecologists (n = 45) used cluster analysis at comparable levels (17.8%) but ordination methods more frequently (51.1%) than stream fish ecologists. Fish community ecologists used ordination methods twice or thrice more frequently (35% of all the fish community studies) than cluster analysis (13%), even though both approaches are used to summarize patterns in fish communities. We suggest that, in part, fish community ecologists use clustering less frequently than ordination methods because researchers are less certain about the degree of interpretability of the solutions when using cluster analysis (i.e., they are unsure about whether the results represent strong or weak patterns, how many clusters should be considered to be present, and which parts of the dendrogram or tree show more reliable information).

It is therefore important that stream fish ecologists be aware of the advances that have occurred in assessing cluster analyses so they can feel comfortable with their use of clustering, be more confident in the interpretation of their results, and better able to distinguish meaningful relationships from more random ones. Stream ecologists may have some familiarity with variants of these methods through their use in phylogenetics or bioinformatics studies (e.g., Eisen et al. 1998; Kerr and Churchill 2001), but generally, most will have little to no familiarity with most or all of these approaches. Our goal is to introduce these methods, demonstrate their use with a common data set of Ontario stream fishes from two river systems, and identify some of the relative merits or shortcomings of these approaches in determining the number of clusters present, the reliability of the clusters, and measures quantifying the association of individual observations to the various clusters. Given the nature of our audience and our goal of making potential users aware of these approaches, we focus on introducing these methods in an ecological setting rather than attempting a more extensive comparison of them through data simulations. Using these tools, all readily available as libraries within the R software

(R Development Core Team 2008), ecologists may be better prepared to apply cluster analysis independently or may also use them in conjunction with ordination approaches to better display and understand data relationships. We introduce the methods using fish community data, both species presence–absence and abundance, but the methods are equally well suited to analyze data based on environmental variables, ecological traits, morphology, or virtually any other set of data.

Fish Community Data

Fish community data from each of two rivers from southern Ontario are used to demonstrate the approaches in cluster analysis. The first river system sampled was the Sydenham River flowing into Lake St. Clair and the second river system was Wilmot Creek, a smaller system flowing into Lake Ontario. Twenty locations in each system were used to provide a balanced design rather than having one system potentially dominating the various analyses. Sampling sites ranged in average width from 1.2 to 11.3 m for Wilmot Creek and from 3.8 to 45 m for the Sydenham River, with a stream length sampled of at least five times the average width. Sites also ranged in local habitat characteristics. Wilmot Creek sites were characterized by a pebble-dominated substrate (55%), meandering stream, with generally intact mixed-wood riparian zone (mean 62.7%). In contrast, the Sydenham River sites were characterized by clay-dominated substrate (72%), largely channelized stream, with overall low deciduous forest cover (22%).

Sampling was conducted using the Ontario Stream Assessment Protocol (OMNR 2007). For this, systematic single-pass electrofishing (at 200 V, 60 Hz, 3 ms) was used at a rate of 5 m/s (OMNR 2007). Block nets were employed upstream and downstream to prevent fish movement out of the sites. All fish were retained in containers of water, identified to species, measured for length and weight, and released. The Sydenham River (Appendix A) had a more diverse, mostly warmwater fish community with 55 species sampled and averaging approximately 17.7 species per site, whereas Wilmot Creek was a cold- to coolwater, salmonid-dominated system having 18 species sampled and averaged only 6.1 species per site. We transformed the species abundance data ($\log x + 1$) given that abundance values ranged over three orders of magnitude and we wanted abundant and less abundant species to have more similar influence in our comparison of sampling sites.

The two systems differed considerably with only 5 species in common out of 68 species found in one or both systems. Therefore, we expected differences would be evident in the various clustering solutions examined, and these differences were an objective in selecting these contrasting fish communities (i.e., this example provides a comparison where we would expect at least two strong groups in the data set). As a means of demonstrating such differences and the degree of variability within the group of sites from each system, we carried out a correspondence analysis (CA) of the fish community data and plotted the observations (i.e., sampling sites) on the first two axes from this ordination (Figure 2). The first axis (20.3% of the total variation summarized) shows a strong contrast between the Sydenham River sites clustered closely on the left end of the first axis and the Wilmot Creek sites positioned on the right-hand end. The Sydenham sites are closely grouped on the second axis (10.2% of the variation), whereas the Wilmot Creek sites show much greater spread encompassing what may be either one or two groups of observations, plus one additional observation positioned at the bottom of the plot—a point

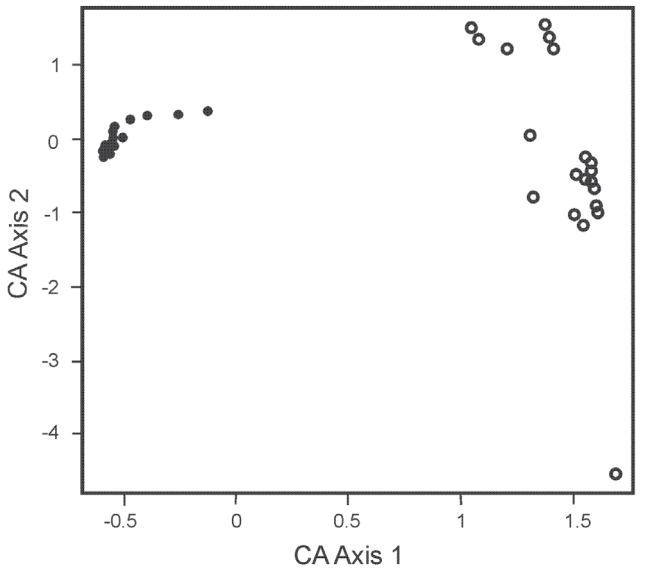

Figure 2. First two axes from a correspondence analysis of the fish community abundance data from 20 sites from Sydenham River (closed circles) and 20 sites from Wilmot Creek (open circles), to show the difference in observations from these river systems.

clearly different from the other observations, suggesting quite different species composition at that sampling site.

Short Review of Clustering Methods

Although our focus is on methods to assess the results of cluster analyses and not on the clustering methods themselves, it is necessary to briefly introduce the clustering approaches that we will use as examples. Not all of these clustering methods can be combined with all of the assessment methods that we review, either because of methodological incompatibility or current lack of available software; however, such issues will be discussed below for each assessment method.

As with most statistical approaches, cluster analyses come in a wide variety of flavors. There is not room to give an exhaustive survey of the various types of clustering here (see for example Legendre and Legendre 1998; Jain et al. 1999; Podani 2000 for such a survey); however, it is helpful to consider a distinction that separates cluster analyses into two broad classes: hierarchical and nonhierarchical. Put simply, hierarchical analyses lead to dendrograms (or "trees"). The branching structure of these trees indicates which objects (e.g., species; sites) are more similar to each other. In contrast, nonhierarchical analyses simply allocate objects into a defined number of groups such that objects within groups are more similar to each other relative to objects between groups. In this paper, we consider one hierarchical approach—unweighted paired group method with arithmetic means (UPGMA)—and three nonhierarchical approaches—*k*-means, *c*-means and linear grouping analysis (LGA)—as representatives of these two major classes of methods.

UPGMA

The first step in UPGMA clustering is the calculation of a resemblance matrix. Such a matrix gives a measure of similarity between each

of the pairs of objects under consideration. The choice of resemblance measure may be of great importance (see Jackson et al. 1989; Legendre and Legendre 1998; Poos et al. 2009 for critiques and advice on how to select a measure appropriate for a given data set). Here we use the chord distance when clustering sampling locations and species (as recommended by Legendre and Legendre 1998; Hirst and Jackson 2007) and which is equivalent to the Euclidean distance on the standardized species abundances (Jackson 1993). The UPGMA algorithm may then be used to construct a tree such that similar pairs of objects are clustered together in the hierarchical branching structure. UPGMA is just one of many approaches for converting a resemblance matrix into a tree; we use it here because it provides a balanced solution to the conflicting problems of space contraction (or chaining) and space expansion. Note that the different choices in the form of hierarchical clustering (e.g., nearest neighbor, furthest neighbor, or UPGMA) can contribute to substantial differences in the branch lengths within and between clusters, which can influence the visual interpretation of the tree (see Legendre and Legendre 1998 for more detail on these issues). Furthest neighbor (complete linkage) clustering makes groups appear to be more different, whereas nearest neighbor (single linkage) tends to minimize or mask differences between groups. These differences are particularly critical when interpreting visual presentations of hierarchical cluster solutions.

K-Means

K-means clustering contrasts with the UPGMA approach in that it is a nonhierarchical method (i.e. it results in only the number of groups, *k*, identified a priori by the researcher. The groups are not nested, but unique and nonoverlapping. However, as with the UPGMA approach, each object to be clustered belongs to only a single group (Legendre and Legendre 1998). *K*-means is based on a criterion to minimize the within-group sum of squares (Hartigan and Wong 1979). With that criterion, the approach is generally based on using the Euclidean-distance resemblance measure, although occasionally, the use of Mahalanobis distance may be used. As Euclidean distance is known to be prone to misrepresenting resemblances of observations based on species abundance (e.g., two sites having no species in common can be shown to be very similar to one another with Euclidean distance simply due to both having low species abundance), *K*-means may not be suitable for use with the original species data (Legendre and Legendre 1998). However, one can summarize such data through the use of ordination methods, resulting in a few axes summarizing major patterns in the communities; therefore, it provides a good companion approach with ordinations for community ecologists. For our example, we use the first two axes resulting from our correspondence analysis, as they provide a simple graphical example consistent with the preprocessing of stream fish community data.

K-means requires an initial estimate at group memberships (often based on random assignment), and then these memberships are refined so as to reduce the within-group sum of squares. It is possible that different initial configurations may produce different results, but most implementations of the approach address this concern through multiple starting points and picking the one that optimizes the sum of squares criterion (Legendre and Legendre 1998).

C-means

C-means is conceptually similar to *K*-means with one important difference (Rousseeuw 1987). *K*-means assumes that an object can belong to only one group even if it shares char-

acteristics of two or more groups; *C*-means relaxes this constraint and allows an object to have affinities to one or more groups and is identified as having its strongest association to one of the groups. Having this fuzzy condition of membership can provide important insight into whether an object might be viewed as having a very strong association to one group or might share characteristics of two or more groups.

In considering stream fish communities, we may have an assemblage at the headwater sites that comprises stenothermal cold- or coolwater fish species and another assemblage of warmwater species found towards the mouth of the river. These two groups of species may have distinct differences in their abundance across the sites, but it is also possible that there may be some species exhibiting a broad distribution across thermal regimes (i.e., eurythermal species) and they are found at most sites. In the *K*-means analysis, these species would need to be placed into either the coldwater or the warmwater assemblage in a two-group *K*-means analysis, even though the eurythermal species would not readily match either of the other two groups of species. Alternatively, in the *C*-means, clustering the eurythermal species could show an affinity to both groups of species, thereby providing a more meaningful ecological representation of the association of the eurythermal species to the other two groups. This temperature-related assemblage structure is a simple example of how nonexclusive group membership might provide insight into how species relate to different habitat characteristics.

Linear Grouping Analysis

Linear grouping analysis (LGA) is also similar to *K*-means (Van Aelst et al. 2006). Like *K*-means, LGA considers an a priori defined number of groups, *k*, such that each observation belongs to only one group. Unlike *K*-means, LGA fits a linear-regression model (i.e., model II regression) to the observations in each group (one model for each group). In our case, we will fit this regression model to the first two correspondence analysis axes of the fish community data. Group membership is chosen to minimize the sum of the squared Euclidean distances of the observations from this linear model. Like *K*-means, this minimizing set of group memberships is usually approximated by repeatedly and randomly dividing the data into *k* groups in order to find the optimal solution, or at least approximate it for large data sets.

Methods Evaluating Inferences about the Cluster Structure of Fish Communities

Here we describe a number of methods for assessing the results of cluster analysis and apply these methods to cluster analyses of our stream fish community data. Table 1 provides a quick summary and comparison of these methods.

Bootstrapping

Within the phylogenetic literature, there is a rich history of developing methods to evaluate trees (i.e., the results of hierarchical cluster analyses). One commonly used approach incorporates resampling theory using the bootstrap (Felsenstein 1985; Efron et al. 1996; Kerr and Churchill 2001; Tonidandel and Overall 2004). Because biological data sets are rarely exhaustive, in the sense that all possible samples have been taken, inferences might be sensitive to the idiosyncrasies of the particular data set that was actually collected. The bootstrap approach to tree evaluation is used to assess the likelihood of this possibility. The general idea is to resample the data without replacement, calculate a tree based on the resampled data,

repeat numerous times, and determine the consistency with which particular branches of the tree occur from each resampled set of data relative to the tree calculated using the original data set.

Various measures have been developed in phylogenetic studies to assess this consistency and the bootstrap probability (BP) is one of the most commonly used (Felsenstein 1985). This approach, like many others, provides a measure that the observed tree is not an artifact of the idiosyncrasies of the particular data that happened to be sampled, but rather a robust conclusion that would have been reached had different data been collected from the same statistical population(s). Index values approaching one for a branch within the tree indicate that the joining (or fusion) of the two components occurs in almost all trees based on the resampled data (i.e., there is strong support for this tree or part of this tree). Such a value is calculated for each branch within a tree. This measure and many others have been shown to be strongly biased, and Shimodaira (2002) provided an alternative called the alternative unbiased (AU) index, which he showed to have superior characteristics. The same bootstrap resampling procedure is used to estimate the AU as in the BP, but the underlying calculation of the AU index differs somewhat. We refer readers to Shimodaira (2002) for details on its derivation and performance characteristics. The interpretation of AU statistics is identical to that for BP statistics; high values of AU indicate consistent groupings, and often values $\geq$ 0.95 are used as the cut-off criterion for a group to show sufficient fidelity to be meaningfully different from a more random association.

We clustered the sampling locations using UPGMA, based on the chord distance. To evaluate the reliability of the resulting tree, we implemented the bootstrapped resampling algorithm found in the R library *pvclust* (Suzuki and Shimodaira 2009). The bootstrapped tree solution shows a strong grouping of the Wilmot Creek sites (Figure 3). All 20 sites from that system are grouped together. The final node in this grouping has an AU = 0.99, indicating that essentially all resampled data sets group the Wilmot sites together rather than including any Sydenham sites within this grouping. Therefore, the Wilmot sites are more similar to one another in terms of their fish community composition than they are to any of the Sydenham sites, and we have a means of providing statistical support to this differentiation. Within the Sydenham River sites, there is a division with one group (comprising sites S2, S7, S8, S16, S17, and S18) being consistently grouped together (AU = 0.95), but no other groups of Sydenham River sites were sufficiently consistent in their grouping (i.e., having a value of AU $\geq$ 0.95) to provide a statistically identifiable group.

In a similar manner, we determined the species associations using UPGMA, based on the Euclidean distance matrix (e.g., Burcher et al. 2008) of the standardized species abundances (i.e., the chord distance). The resulting tree (Figure 4) shows a series of clusters identified as being meaningful (i.e., AU values of 0.95 or greater). The first "significant" cluster and the one showing the greatest overall similarity, contains longnose dace *Rhinichthys cataractae* and rainbow darter *Etheostoma caeruleum*, both species being absent from all Sydenham sites but present in many Wilmot sites (Appendix A). The second most strongly associated cluster comprises warmwater species such as largemouth bass *Micropterus salmoides* and bluegill *Lepomis macrochirus*, species found in a few Sydenham sites but no Wilmot sites. A group of similar strength clustered ghost shiner *Notropis buchanani* and gizzard shad *Dorosoma cepedianum*, again based on their presence in only a few Sydenham sites. A cluster

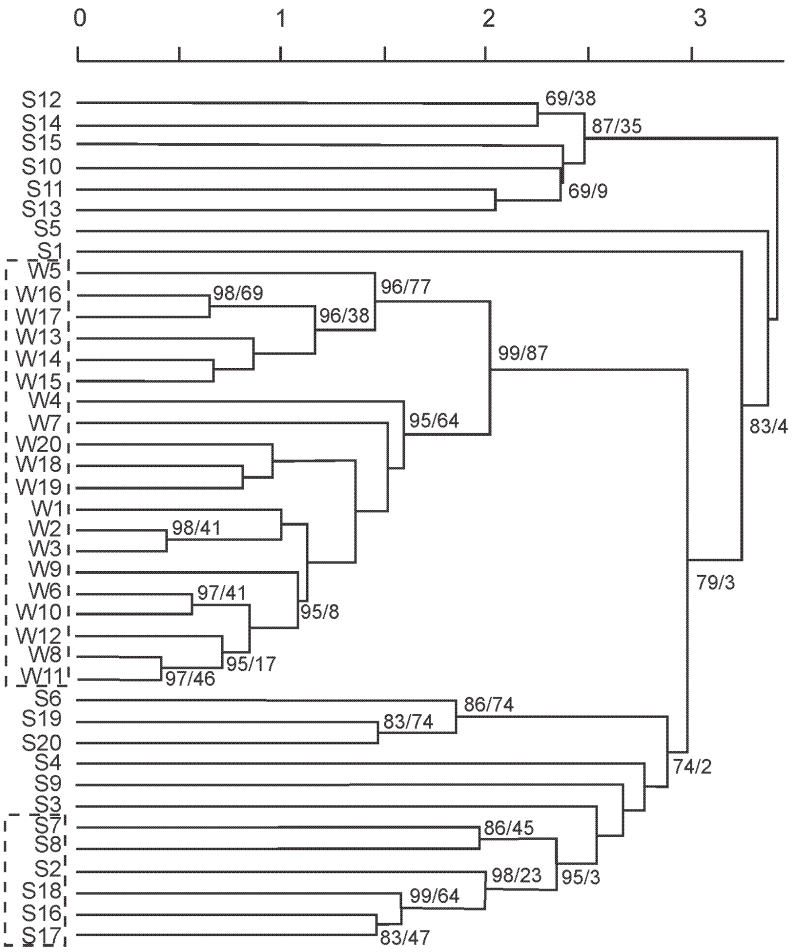

Figure 3. Unweighted paired group method with arithmetic means dendrogram of 40 sites sampled from two rivers in Ontario. The clustering was based on the Euclidean distance measure of standardized fish community abundance data (i.e., chord distance). Twenty sites from Sydenham River are denoted as S1–S20 and sites from Wilmot Creek are identified as W1–W20. In the couplet of values presented at each node, the first value is the approximately unbiased (AU) estimate and the second one is the bootstrapped probability (BP) estimate. Note that not all nodes include the AU/BP estimates, simply to enhance visual clarity and legibility.

comprising coldwater species such as rainbow trout *Oncorhynchus mykiss* and mottled sculpin *Cottus bairdii* was due to their common abundance at most sampling locations in Wilmot Creek, but their absence from Sydenham River sites. Such strong associations of species assemblages within a river system and their absence from the other system contributed to the separation of river sites in Figures 2 and 3. As the Sydenham River sites show great variation in their assemblage structure and a less defined resemblance of their sites based on these fishes, the sites are not closely grouped in either Figures 2 or 3. Sydenham River sites show the occurrence of species such as mimic shiner *N. volucellus*, shorthead redhorse *Moxostoma macrolepidotum*, longear sunfish *L. megalotis*, fantail darter *E. flabellare*, and greater redhorse

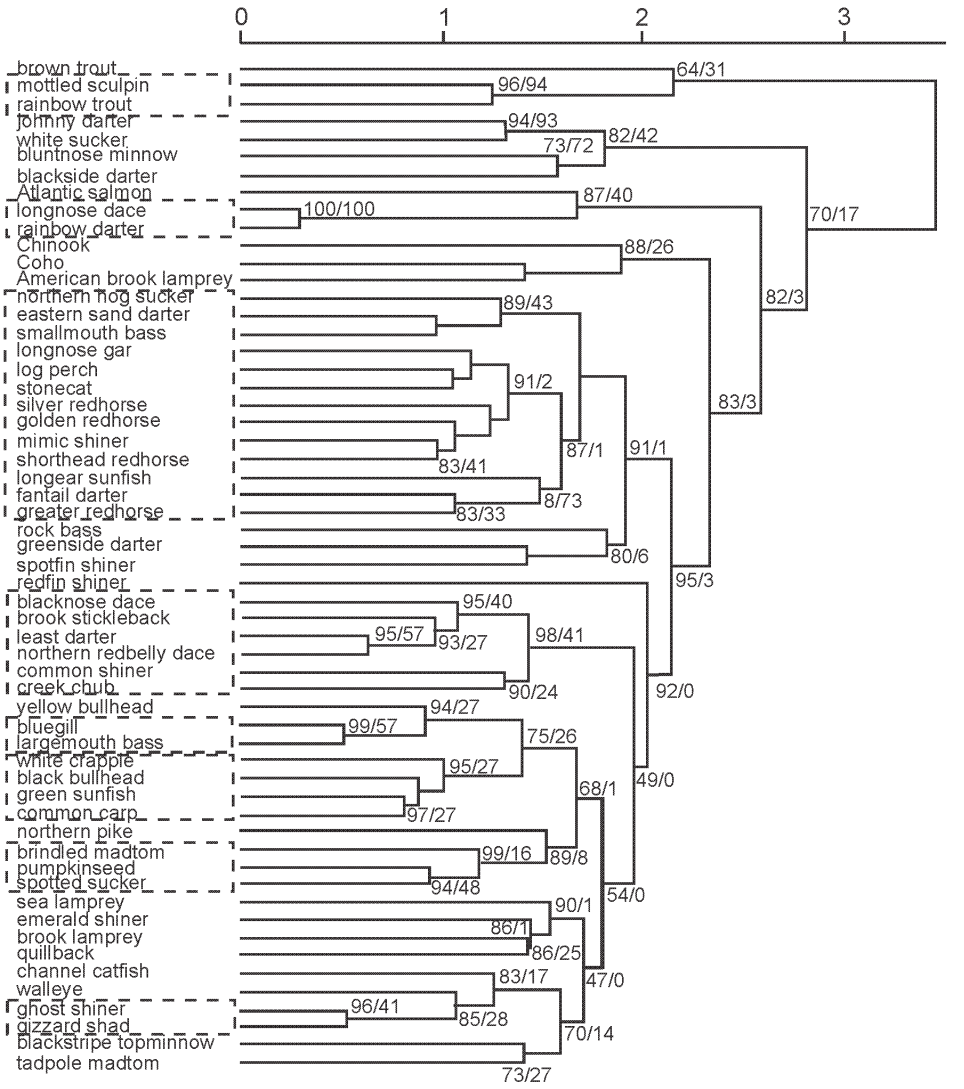

Figure 4. Unweighted paired group method with arithmetic means dendrogram of fish species from 40 sites sampled from two rivers in Ontario. The clustering was based on the Euclidean distance measure of fish community abundance data standardized to range between 0 and 1 for each species. In the couplet of values presented at each node, the first value is the approximately unbiased (AU) estimate and the second one is the bootstrapped probability (BP) estimate. Note that not all nodes include the AU/BP estimates, simply to enhance visual clarity and legibility.

M. valenciennesi, at various locations within the set of sampled locations, but their absence from other Sydenham locations. Similarly, the final two assemblages contain white crappie *Pomoxis annularis*, black bullhead *Ameiurus melas*, green sunfish *L. cyanellus*, and common carp *Cyprinus carpio* in the first and brindled madtom *Noturus miurus*, pumpkinseed *L. gibbosus*, and spotted sucker *Minytrema melanops* in the second. These two assemblages again contrast a series of sites within the Sydenham from all other sites, and generally the presence of the first set of species is matched with the absence of the second set (and vice versa)

within these locations. The species in these last four assemblages are all absent from all Wilmot Creek sites.

Through the clusters provided by these two bootstrapped dendrograms, we can define assemblages that are associated with particular sets of sampling locations. In various instances, researchers use ordination methods to determine the association between their variables and observations, but we can achieve similar goals with a bootstrapped cluster analysis and also determine which sets of sites or species represent statistically strong signals (i.e. sites having very similar species assemblages and sets of species representing strongly repeatable assemblages). In many instances, ecologists are seeking to determine species assemblages that are repeatable in order to better understand the mechanisms controlling species distributions and those species that may have strong interactions. Identifying groups of sites containing similar species is an important step in determining which factors or mechanisms may be leading to community composition and maintenance, be those factors historical biogeographic ones or current abiotic and biotic conditions (Jackson et al. 2001).

Nemec and Brinkhurst (1988) first introduced the bootstrap methodology for clustering analysis to ecologists, although their method requires replicate observations for each sampling location. Pillar (1999) provided a more generalized application of the bootstrap approach to cluster analysis in ecology, with examples based on both simulated and plant vegetation communities. McKenna et al. (2008) provided one of the rare examples of a bootstrapped cluster analysis being used in community ecology in their study of Lake Erie ichthyoplankton. Although it has considerable promise, the bootstrapping approach has been largely ignored by ecologists in general. It is one of the few measures that allows a more formal evaluation of the fidelity of individual clusters and does not require that all observations be included in the final clusters that are interpreted as being meaningful, nor that the number of groups need be defined a priori (Table 1). One technical caveat we note is that the underlying objects to be resampled (e.g., the fish species are resampled when bootstrapping the sampling location data set) are assumed to be independent. Given that it is unlikely that the abundances of various species within a local assemblage are independent of one another, one should consider the bootstrap approach to provide a relative measure of the degree of association rather than perhaps providing an estimate of the underlying statistical probability related to testing a null hypothesis.

Calinski and Simple Structure Indices

Choosing an appropriate k number of groups is important to successful nonhierarchical clustering, as determining the number of groups is the necessary and first step in determining which observations are grouped together. Therefore, as fish community ecologists, we need a suitable tool to aid us in making such assessments or we risk interpreting our data either less effectively or simply incorrectly. Two approaches can be followed in defining the number of groups. The first approach is based on the instance where the researcher has some fundamental reason for defining k groups (e.g., for some specific reason, the researcher wants to divide the observations into three groups only). As this approach is based on some rationale defined by the researcher rather than a statistical approach, we will not consider its deliberation further. The other approach, and our focus here, is to calculate a measure of how good is the solution based on k groups and compare it to the solutions based on more or less groups being defined.

We consider two indices used to compare different choices for k in a K-means cluster

analysis, using the *cascadeKM* function in the R library *vegan* (Oksanen et al. 2009). The first measure considered is the Calinski index, which is essentially the analysis of variance statistic comparing the sum of squares among groups or clusters relative to the within-group sum of squares. The second measure is the simple structure index (SSI), which provides a measure integrating the maximum difference for each variable between clusters, how different the centroid values are for each variable relative to the overall variable mean, and a measure of the size of the most contrasting cluster. In order to use these measures, we calculate either index when defining two groups and recalculate the indices for three groups, four groups, and so forth. Where the index is maximized defines the number of groups providing the selected *k*-means solution.

The Calinski index proved to be a difficult one to reconcile with results in our data analyses; this index selected $k = 19$ groups, which is clearly unreasonable (see Figure 5 and Table 1). Tibshirani et al. (2001) found similar problems with the Calinski index in some of their simulation studies, particularly when variables were strongly correlated within groups. Another potential issue with the Calinski index is that it is expected to perform best when the clusters are relatively equal in size; in fact, it is suggested to be the best measure in these cases (Milligan and Morgan 1985; R manual for Vegan library function *cascadeKM*; Oksanen et al. 2009). Although two clusters are identical in size when only two clusters are considered (i.e., each containing the 20 sites from one of the two rivers), neither measure suggests that this result is the optimal division. For the case of three clusters (Figure 6), the Sydenham River sites are all grouped together and the Wilmot Creek sites comprise two groups, with one containing 6 sites (W5 and W13–W17) and the third cluster containing the remaining 14 Wilmot sites—clearly presenting an imbalance in the

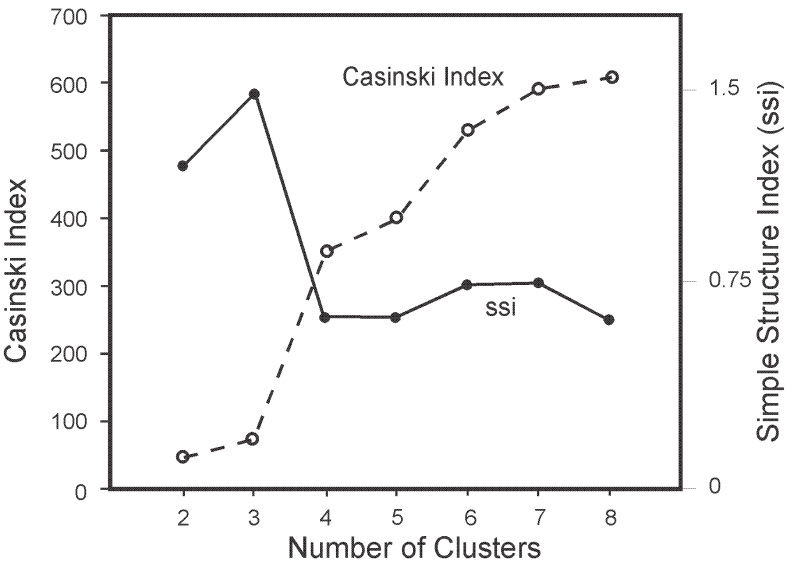

Figure 5. Two indices associated with the *K*-means cluster analysis solution relative to the number of clusters. For both indices, the peak in the index value indicates the optimal solution. Note that the Calinski index continues to increase as the number of clusters increases, indicating that a suitable solution was not suggested by this measure.

Table 1. Summary and comparison of the cluster analysis assessment methods reviewed in this study. LGA = linear grouping analysis

	Bootstrap	Calinski index	Simple structure index	Gap analysis	Silhouette plots
Hierarchical assessment?	Yes	No	No	No	No
Choose number of groups?	Yes	Yes	Yes	Yes	Yes
Assesses affinities of object to clusters?	Yes	No	No	No	Yes
R package used	*pvclust*	*vegan*	*vegan*	*lga*	*cluster*
Performance with our stream fish data	Allowed us to statistically assess which species assemblages were associated with which site clusters.	Selected an unreasonably large number (19) of groups.	Selected three groups, which matches intuition.	Could not assess performance because gap analysis is available in R only for assessing LGA results, which were not appropriate for our data.	Selected two or three groups, which matches intuition.

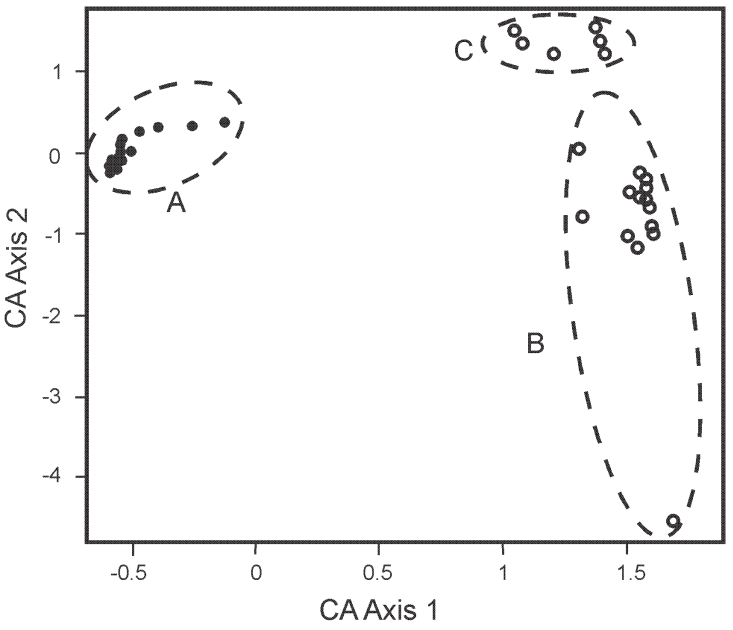

Figure 6. Correspondence analysis plot showing the three groups defined by *K*-means. The number of groups, $k = 3$, was selected by the simple structure index (SSI).

size of the various clusters. Therefore, researchers are cautioned that in cases where the size of their clusters differs, likely to be a commonly encountered case for stream fish ecologists, the Calinski index may not provide an ideal measure. By contrast, the SSI measure indicates that the three-cluster solution (shown in Figure 6) is best, which is intuitively reasonable given the three visually identifiable clusters in the correspondence analysis (Figure 6; Table 1). When working with only two-dimensional ordination solutions as their input data, as we are here, researchers have the advantage of being able to compare visually whether the clusters appear to make sense too, but such opportunities are unavailable when working with higher dimensional data or with the raw species data or habitat data.

We are unaware of any studies that have used either the Calinski or simple structure indices to analyze stream fish communities. However, Sowa et al. (2007) used the Calinski index to analyze environmental data in their work on riverine conservation in Missouri.

Gap Analysis

Gap analysis (Tibshirani et al. 2001) provides another method for selecting the number of groups, *k*. It was developed to address a problem with the following naïve approach to choosing *k*. Many clustering methods, such as *K*-means and linear grouping analysis, determine group membership by minimizing a measure of the dispersion of points within each cluster. One might naively attempt to choose *k* by minimizing such a dispersion measure (e.g., pooled within-group variance or sums of squares). Such dispersion measures generally follow a monotonic decrease as the number of clusters increases, regardless of the cluster structure inherent in the data (i.e., the criterion is minimized when each object forms a different cluster), but the result is obviously less than informative. Hence researchers often determine the point at which this monotonic relationship begins to flatten. This point of flattening is interpreted as the point at which the number of clusters is meaningful and additional clusters would be less informative; this approach is conceptually identical to the scree plot that has been used to determine the number of interpretable ordination dimensions (Jackson 1993; Peres-Neto et al. 2005). The problem that gap analysis seeks to overcome is that, in practice, such break-points are often not found or at least not simple to interpret.

Gap analysis provides a solution to this problem with virtually all clustering methods that (1) are based on a measure of within-cluster dispersion, and (2) require a choice for the number of groups (Tibshirani et al. 2001). Most commonly however, it is applied to methods based on a squared Euclidean distance. In particular, gap analysis is currently only available for assessing the cluster structure inferred from a linear grouping analysis (highlighted in Table 1). Essentially, the method calculates the sum of squared Euclidean distances for all points within a cluster and pools these values from all k clusters to provide the pooled within-cluster sum of squares (denoted as W_k). This observed W_k value is then compared against a null reference distribution of W_k values. This reference distribution assumes that each observation is equally likely to be included in any of the *k* groups (i.e., a uniform distribution). Specifically, the expected value (i.e., mean; E_k) of the reference distribution is calculated by repeatedly drawing multiple random samples, calculating W_k for each sample, and averaging these W_k values. The gap statistic is the difference between the expected and observed values, $\log(E_k) - \log(W_k)$. This statistic is calculated for various numbers of clusters and, where maximized, provides the estimate of the number of clusters in the data.

Gap analysis was used to estimate the number of clusters in a linear grouping analysis of our fish community ordination and implemented with the R library *lga* (Harrington 2008). The plot (Figure 7A) of the logarithm of the pooled within-cluster sum of squares shows the characteristic decline as the number of clusters increases in both the observed data (W_k) and that expected based on multiple samples drawn from the reference distribution (E_k). The plot of the gap statistic relative to the number of clusters (Figure 7B) considered shows a different pattern, peaking at three clusters and then declining, thereby indicating three clusters as being the optimal solution. However, the value for E_k is based on b reference samples (i.e., b is the number of resampled data sets used; e.g., b = 10,000), and a standard error associated with this estimate can be derived also. Figure 7B also shows standard errors associated with the estimates for each number of clusters. As the procedure is based on a resampling approach, it is important that sufficient numbers of reference samples are used in estimating the expected value (Jackson and Somers 1989). For example, in this study, the repeated use of 1,000 reference samples provided different outcomes, which would

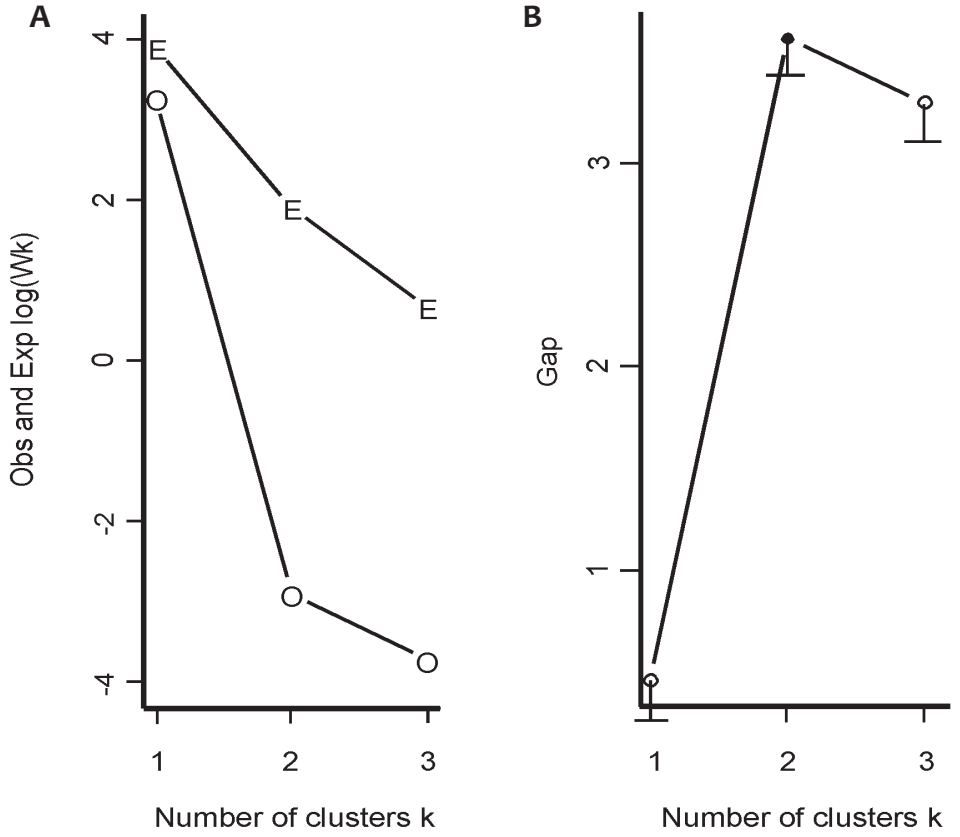

Figure 7. (A) Plot of gap analysis results showing the log-transformed values of the observed and expected statistics versus the number of clusters used in the solution. (B) Plot showing the corresponding gap statistic and standard errors relative to the number of clusters. The statistic peaks with three clusters, but there is considerable overlap in the standard errors between the three and four cluster solutions.

have lead to anywhere from two to four cluster solutions being concluded as being the "best" solution. However, the use of a greater number of simulations (i.e., 10,000 in this case) resulted in a stable outcome of a two-cluster solution (Figure 7).

Although Tibshirani et al. (2001) indicated that the gap statistic can be used with virtually any type of clustering, within R it is associated with the library *lga*, which provides a very specific form of cluster analysis. Having defined $k = 2$ as being the best solution from the gap analysis, we determined the group assignments, and these are shown in Figure 8. From the results shown, we can see that the assignment of observations to the different groups does not make intuitive sense and differs considerably from the results obtained from the other clustering solutions. Three of the Wilmot Creek sampling sites have been grouped (open circles in upper right region of Figure 8) with the set of Sydenham sites. Although this group best fits the criterion for the linear group analysis, the community data being analyzed are not well handled through a linear model (highlighted in Table 1). We have used a two-dimensional summary of the community data by using the first two axes from a correspondence analysis. There is no reason to assume that linear relationships within an ordination plot should result in linear relationships between axes for individual groups of observations. Similarly, given that community data are frequently rich in zero abundance values, this linear model is not likely to provide a good summary of group structure within the original abundance data.

For general use with community data, we do not advocate the use of the LGA approach given the underlying relationships in such data and the problems posed by imposing multiple linear relationships to fit to the

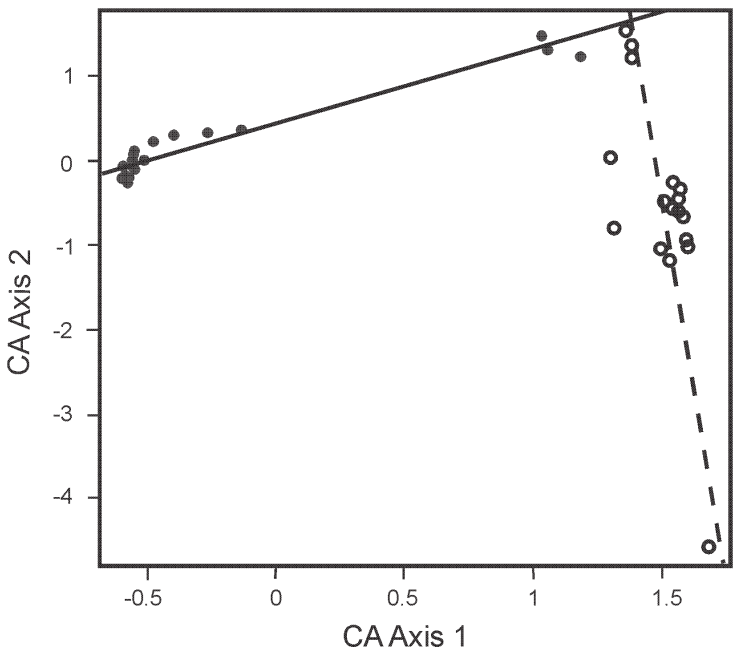

Figure 8. CA plot showing results of gap analysis and linear group analysis defining a two-group solution. Note that three Wilmot Creek sites (open circles) have been grouped with the Sydenham River sites (closed circles).

data. For example, if stream habitat data were being analyzed to determine relationships and groupings of sampling locations, the linear relationship between some habitat variables may provide excellent data for this methodology. Given that different ecoregions or smaller units may differ in the linear relationships (i.e., slope, intercept) between stream habitat variables, the linear grouping analysis approach would be well suited to identify these relationships and therefore group these sites. Similarly, if a researcher is focused on only a few species that tend to be encountered at most sites, this method may hold promise too. However, the extension of gap analysis to other clustering approaches may provide interesting opportunities for further methodological research, and it is to be encouraged.

Silhouette Plots

Silhouette plots (Rousseeuw 1987) provide a graphical and quantitative approach to determine the optimal number of clusters. Silhouette plots can be used to identify the strength of clusters individually and collectively, as well as how well each observation matches its cluster assignment. The name is derived from the plot's resemblance to a silhouette of a city where each object clustered represents the equivalent of a building in the silhouette. Silhouette plots are typically used with nonhierarchical clustering approaches, and these can include techniques such as *K*-means or *c*-means. The approach works with clusters containing two objects or more (i.e., it cannot be used to evaluate singleton observations). The underlying premise is to measure the association between object *i* (a sampling location in this case) with other objects from the same group or cluster and the strength of this association to object i's relationship to objects in other clusters. The associated measure, $s(i)$, ranges from –1 to +1, where a value close to +1 indicates that object *i* is strongly associated with other members of the cluster in which it is grouped. As $s(i)$ decreases and approaches 0, it indicates that object *i* is not well associated with any particular cluster. Values less than 0 indicate that object *i* is actually better suited for inclusion within a different cluster than the one in which it has been placed. It may seem counterintuitive that a cluster analysis could define groups in which members are not well suited for inclusion within their respective group but are in fact better suited for inclusion into another group. Cluster analysis (and agglomerative, hierarchical clustering in particular) is known to result in such outcomes depending upon the characteristics of the data and the type of resemblance measure and clustering algorithm used (Jackson et al. 1989; Legendre and Legendre 1998; Poos et al. 2009), perhaps leading researchers to be somewhat wary of cluster analysis in general. However, approaches such as silhouette plots allow us to better determine whether such groupings are informative or not.

As the index $s(i)$ is calculated for each object, we can determine the range of values obtained for all objects within a group, an average for each group, and an overall average value for the entire tree. Results from a cluster analysis can be examined for group structure ranging from two groups and up. Generally, one will try different numbers of groups in order to determine how the resulting group average values and overall tree indices change in response. The number of groups leading to where the overall silhouette average is maximized provides a means of estimating how many groups of objects should be considered.

Silhouette plots can be used when clustering is done on the original data set or using variables representing multivariate axes. We use the site scores from the first two correspondence analysis axes, as this case provides an example that can be considered

easily in terms of its graphical nature (Figure 9). Various distance measures can be used to quantify how each object relates to all others, and we have used Euclidean distance again for this example as it provides a simple geometric measure of resemblance (i.e., our straight-line distance measure) and is appropriate given that we are quantifying distances in ordination space. We then carried out a series of *C*-means cluster analyses, sequentially increasing the number of groups in each analysis in order to determine which solution provides the best outcome based on the average silhouette index and therefore how many groups we should interpret. The *C*-means clustering and silhouette plots were calculated using the fanny function in the R library *cluster* (Maechler et al. 2009).

As an example, we show the results obtained from the three-group solution: Group A for the Sydenham River and Groups B and C for Wilmot Creek. We see that in Group A, almost all Sydenham sites have silhouette widths, $s(i)$ exceeding 0.80 (Table 2; Figure 9), the value recommended to suggest strong membership (Rousseeuw 1987). Two Sydenham sites, S19 and S20, had values between 0.60 and 0.80, which suggests good agreement with the other group members. Therefore, the Sydenham River sites all tend to form a strong cluster, and this is supported by the average silhouette width being 0.89 for this group. The Wilmot Creek sites were divided into two clusters: B and C. Cluster B had an average silhouette width of 0.60 and ranged from 0.19 to 0.71, suggesting that this final grouping is weakly defined relative to the other two clusters. The remaining Wilmot sites clustered into the final group, cluster C, also showed very strong group membership and comprised W5 and W13–W17 sites with an average silhouette width of 0.87.

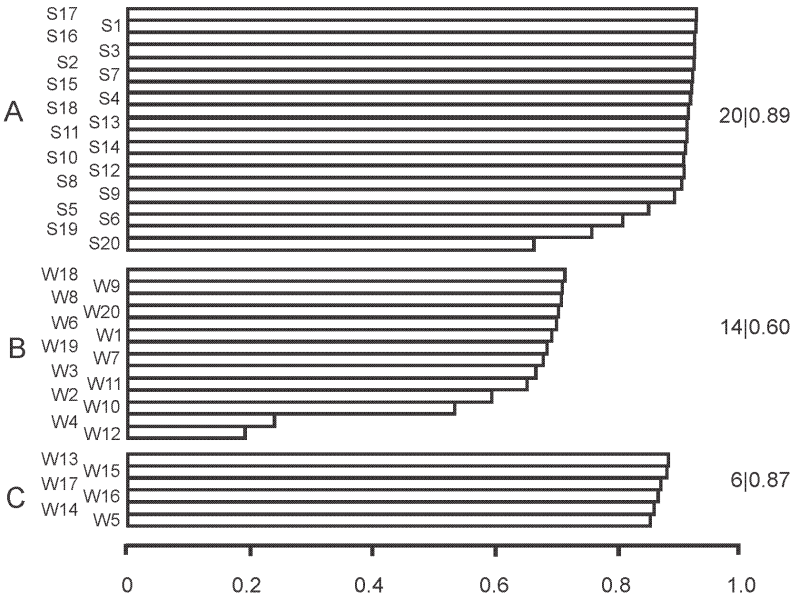

Figure 9. Silhouette plot results for a three-group solution. Each bar represents one observation, and the longer the bar, the stronger the association between that observation and its cluster. The numbers to the right of the bars represent the number of observations within the cluster and the average silhouette index for that group (e.g., Group A has 20 sites and an average silhouette value of 0.89, indicating a strongly defined cluster). Values greater than 0.8 indicate a strong affinity of the group members to resemble one another based on their species composition.

Table 2. Results from the *C*-means clustering and silhouette analysis of the 40 sampling locations. The silhouette width [*S*(*i*)] provides a measure of how strong the affinity is between an individual observation and the group to which it has been assigned. Values range from −1 to +1, and values exceeding 0.80 indicate a strong affinity. The probability of membership for each sampling location into each of the three defined groups is shown under the three group headings. All locations from Sydenham River (S1–S20) are grouped together, whereas the sites from Wilmot Creek (W prefix) are clustered into two groups.

Sampling site	*S*(*i*)	Group A	Group B	Group C	Sampling site	*S*(*i*)	Group A	Group B	Group C
S1	0.93	100	0	0	W1	0.69	0	99	0
S2	0.93	100	0	0	W2	0.59	1	98	2
S3	0.93	100	0	0	W3	0.67	0	99	1
S4	0.92	100	0	0	W4	0.24	26	53	21
S5	0.85	98	1	1	W5	0.85	1	1	99
S6	0.81	97	2	2	W6	0.70	0	100	0
S7	0.92	100	0	0	W7	0.68	1	98	1
S8	0.90	100	0	0	W8	0.71	0	100	0
S9	0.89	99	0	0	W9	0.71	1	98	1
S10	0.91	100	0	0	W10	0.53	1	96	3
S11	0.91	100	0	0	W11	0.65	0	99	1
S12	0.91	100	0	0	W12	0.19	7	75	18
S13	0.91	100	0	0	W13	0.88	0	0	100
S14	0.91	100	0	0	W14	0.86	0	1	99
S15	0.92	100	0	0	W15	0.88	0	0	100
S16	0.93	100	0	0	W16	0.87	0	0	99
S17	0.93	100	0	0	W17	0.87	0	1	99
S18	0.92	100	0	0	W18	0.71	1	99	1
S19	0.76	94	3	4	W19	0.68	2	96	2
S20	0.66	87	5	7	W20	0.71	1	98	1

Examining the placement of observations into the groups shows similar information, as can be inferred between the groups. For example, W4 shows a low silhouette width (0.24) and was placed into cluster B, but only with a probability of 0.53 meaning that there was almost an equal likelihood of either belonging in cluster B or not belonging in cluster B. Such results allow us to view clusters A and C as being well defined and would be good candidates for detailed interpretation of their fish communities and having different assemblage being present in each of the two groups. However, a researcher would want to demonstrate much greater caution in considering cluster B as representing a well-defined group of sampling locations as they are not characterized by a similar set of fish species across these sites.

We are not aware of silhouette analysis being used in a study of fish communities, stream or otherwise; however, Schaefer and Wilson (2002) provided an example of its use in the genetic structure of walleye populations in Lake Erie. Again, this method has clear promise in allowing researchers to better identify the fidelity of observations to groups (clusters) and the relative strength of those groups, once the issue of the number of groups has been resolved.

Comparability of Defined Clusters

The Sydenham River and Wilmot Creek represent fish assemblages from two completely different systems in Ontario, Canada. These systems differed not just in their locality, but also in the types of local habitat conditions (e.g., pebble/gravel substrates versus clay),

thermal regime (cool versus warm), and ecozone (mixed deciduous–coniferous versus deciduous forest). In total, only five species were found in both systems. As such, we would expect that all methods would produce clusters that clearly segregated the fish assemblages between those two areas.

Examining the results of the interpretable clusters defined by the various methods provides interesting comparisons (Table 3). The approach based on bootstrapping provides a hierarchical clustering of all observations (or species), but not all members of the tree will necessarily represent "significant" groups

Table 3. Results of the cluster membership defined by each of the four methods considered in the study. Note that some of the methods group all observations, whereas other methods (e.g. bootstrapping) may selectively included observations and not retain others in clusters. SSI = simple structure index; LGA = linear grouping analysis.

	Bootstrapping	SSI/*k*-means	Gap analysis/LGA	Silhouette plots
S1	–	A	A	A
S2	A	A	A	A
S3	–	A	A	A
S4	–	A	A	A
S5	–	A	A	A
S6	–	A	A	A
S7	A	A	A	A
S8	A	A	A	A
S9	–	A	A	A
S10	–	A	A	A
S11	–	A	A	A
S12	–	A	A	A
S13	–	A	A	A
S14	–	A	A	A
S15	–	A	A	A
S16	A	A	A	A
S17	A	A	A	A
S18	A	A	A	A
S19	–	A	A	A
S20	–	A	A	A
W1	B	B	B	B
W2	B	B	B	B
W3	B	B	B	B
W4	B	B	B	B
W5	B	C	A	C
W6	B	B	B	B
W7	B	B	B	B
W8	B	B	B	B
W9	B	B	B	B
W10	B	B	B	B
W11	B	B	B	B
W12	B	B	B	B
W13	B	C	B	C
W14	B	C	B	C
W15	B	C	A	C
W16	B	C	A	C
W17	B	C	B	C
W18	B	B	B	B
W19	B	B	B	B
W20	B	B	B	B

or clusters. In Table 3, we see a strong group of Wilmot Creek sites in a single cluster, whereas only a subset of the Sydenham sites is grouped together. These groupings differ substantially from those produced by the other three methods and, at first consideration, may suggest that the technique performs differently. However, note that for the purposes of our demonstration, we used the standardized species abundance data in the bootstrap analysis, whereas we used ordination axes as the input to the other methods to better present the findings—this alone may contribute to differences between the bootstrapped solution and the other methods. We include both approaches (some based on the original data and some based on the ordination axes) to highlight the flexibility of methods that can be used, and also to note limitations or problems that may arise when trying to use a method with data types that may not be appropriate or well suited. Furthermore, our use of different forms of input data are also to suggest to the reader that there may be different approaches that can be applied to their data regardless of whether it represents presence–absence data through to continuous, abundance data. Although the bootstrapping method did not group the various Sydenham River sites together, all other methods did and showed them to be a strongly defined cluster, consistent with the results shown in the CA plot (Figure 2). However, these three methods differ in how they defined group membership for Wilmot Creek sites W5 and W13–W17. Where such differences arise may help define observations for which we are less certain in their affinity, but the consistent grouping of most observations provides a much greater degree of confidence in our findings and can help stream fish ecologists better define useful species assemblages and groups of sites sharing similar assemblages.

Summary and Conclusions

Stream community ecologists are generally faced with challenges in summarizing their data through the use of various ordination and clustering approaches. Ecologists have generally incorporated a variety of the assessment tools that have been introduced to help guide in the assessment of ordination solutions, but generally have not incorporated the comparable advances available for cluster analyses. Our goal has been to demonstrate a series of the more promising approaches for identifying site and species groupings. With hierarchical clustering approaches, the use of bootstrapping techniques provide means of quantifying the overall structure and how many clusters should be considered to be interpretable (Table 1). The bootstrap approach provides greater detail about the hierarchical structure of the results. It allows the highlighting of particular areas within a dendrogram that show highly consistent clustering over many bootstrapped outcomes, identifying groups having strong fidelity. Bootstrapped analysis may provide a conservative estimate of what portions of the dendrogram can be interpreted, in particular if the 95% level is used as the cut-off criterion. As a consequence, one may find that large amounts of the dendrogram may be considered as "nonsignificant" (i.e., lacking identity to a particular group) in many solutions—an outcome that may be less satisfying for many researchers but which should serve to caution researchers on potential overinterpretation of their findings.

Both gap analysis and bootstrapping are based on resampling approaches, and it is important that researchers use a large number of resampled data sets in order to derive stable solutions (Jackson and Somers 1989). Given that computational demands generally no longer restrict our abilities to conduct large resa-

mpling exercises, a conservative standard may be to apply 10,000 bootstrapped data sets in the bootstrapped and gap analysis approaches. As the implementation of gap analysis within R appears to be limited currently to its connection with linear group analysis, and given that community data typically do not demonstrate strong linear relationships (e.g., many zero values, nonlinear relationships in abundance between species), it cannot be recommended as a general approach for analyzing fish communities (Table 1). However, it may be a very useful approach to consider with the environmental data associated with community analyses. Once gap analysis becomes more widely available as a diagnostic tool for other clustering solutions, it may provide an interesting and valuable tool for community ecologists.

Silhouette plots provide researchers with different tools to assess both the number of clusters suitable for interpretation and how well suited each observation is matched to each cluster of observations. The approach can be used with either hard or fuzzy clustering, and the fuzzy clustering approach allows the assessment of whether points match strongly with the group in which they are placed. Knowing whether an observation is very well matched to other observations or whether such an observation may be almost equally well suited with another set of observations can provide valuable insight to ecologists. For example, this information would allow a researcher to know how tightly grouped a set of sampling observations may be or, alternatively, whether a particular species of fish is part of a strongly defined assemblage or whether that species may share characteristics with multiple assemblages. Clearly, such insight can aid ecologists in better understanding the strength of the associations that they see in community analyses.

Karr and Martin (1981) once challenged the utility of principal component analysis in community ecology by suggesting that one could interpret results obtained from random data. Although ordination and clustering methods will provide solutions even if the data are random, many studies advancing methods on interpreting ordination results have provided tools that, when applied appropriately, can allow us to easily address the concerns raised by Karr and Martin—we can now dismiss such concerns through the appropriate analysis of our data. Many community ecologists have readily adopted these methods and contributed to their development in many cases. There is also a rich source of approaches that allow stream fish ecologists to move beyond the standard approaches that we have used in cluster analyses. Many of these methods are used commonly in other fields but have been virtually unused by fish ecologists—a situation likely due to researchers not knowing about their existence and attributes. Many ecologists may have avoided cluster analysis due to the less formal ways of determining the strengths of the summarized relationships—we can now move beyond those concerns. These methods provide an excellent addition to our quantitative toolbox and we advocate their use in conjunction with the visual interpretation of the dendrograms. We caution against simply interpreting cluster analysis solutions solely from a visual assessment as we see this as being analogous to simply interpreting whether bivariate data are well correlated solely by looking at a scatterplot and not evaluating the relative strength of the association through quantitative measures such as correlation coefficients—both visual and quantitative approaches work together to enhance our interpretation and provide support in convincing others about underlying ecological structure. All methods we have presented are freely available in R (R Development Core Team 2010), thereby providing fish ecologists

with additional tools to better define patterns in community data. As community ecologists have implemented many approaches to better understand our ordination solutions, it is now time to consider doing the same with our cluster analysis approaches.

Acknowledgments

We gratefully acknowledge the funding provided by NSERC Canada that allowed this work to be conducted. As well, we thank the Ontario Ministry of Natural Resources, and Les Stanfield in particular, for collecting and making available the Wilmot Creek data. We thank Keith Gido and two reviewers for their helpful comments and suggestions on our chapter.

References

Bowman, M. F., R. Ingram, R. A. Reid, K. M. Somers, N. D. Yan, A. M. Paterson, G. E. Morgan, and J. M. Gunn. 2008. Temporal and spatial concordance in community composition of phytoplankton, zooplankton, macroinvertebrate, crayfish, and fish on the Precambrian Shield. Canadian Journal of Fisheries and Aquatic Sciences 65:919–932.

Burcher, C. L., M. E. McTammany, E. F. Benfield, and G. S. Helfman. 2008. Fish assemblage responses to forest cover. Environmental Management 41:336–346.

Efron, B., E. Halloran, and S. Holmes. 1996. Bootstrap confidence levels for phylogenetic trees. Proceedings of the National Academy of Sciences 93:13429–13434.

Eisen, M. B., P. T. Spellman, P. O Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences 95:14863–14868.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791.

Frimpong, E. A., and P. L. Angermeier. 2010. Trait-based approaches in the analysis of stream fish communities. Pages 109–136 *in* K. B. Gido and D. A. Jackson, editors. Community ecology of stream fishes: concepts, approaches, and techniques. American Fisheries Society, Symposium 73, Bethesda, Maryland.

Grossman, G. D., Nickerson, D. M., and M. C. Freeman. 1991. Principal component analysis of assemblage structure data: utility of tests based on eigenvalues. Ecology 72:341–347.

Guisan, A., and N. E. Zimmermann. 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135:147–186.

Harrington, J. 2008. lga: tools for linear grouping analysis. R library

Hartigan, J. A., and M. A. Wong. 1979. A k-means clustering algorithm. Applied Statistics 28:100–108.

Hirst, C. N., and D. A. Jackson. 2007. Reconstructing community relationships: the impact of sampling error, ordination approach and gradient length. Diversity and Distributions 13:361–371.

Higgins, C. L., and G. R. Wilde. 2005. The role of salinity in structuring fish assemblages in a prairie stream system. Hydrobiologia 549:197–203.

Jackson, D. A. 1993. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. Ecology 74:2204–2214.

Jackson, D. A., P. R. Peres-Neto, and J. D. Olden. 2001. What controls who is where in freshwater fish communities: the roles of biotic, abiotic, and spatial factors. Canadian Journal of Fisheries and Aquatic Sciences 58:157–170.

Jackson, D. A., and K. M. Somers. 1989. Are probability estimates from the permutation model of Mantel's test stable? Canadian Journal of Zoology 67:766–769.

Jackson, D. A., K. M. Somers, and H. H. Harvey. 1989. Similarity coefficients: measures of co-occurrence and association or simply measures of occurrence? American Naturalist 133:436–453.

Jain, A. K., M. N. Murty, and P. J. Flynn. 1999. Data clustering: a review. ACM Computing Surveys 31:264–323.

Joergensen, O. A., C. Hvingel, P. R. Moeller, and M. A. Treble. 2005. Identification and mapping of bottom fish assemblages in Davis Strait and southern Baffin Bay. Canadian Journal of Fisheries and Aquatic Sciences 62:1833–1852.

Kaufman, L. and P. J. Rousseeuw. 2005. Finding groups in data: an introduction to cluster analysis. Wiley, New York.

Karr, J. R., and T. E. Martin. 1981. Random numbers and principal components: further searches for the unicorn? Pages 20–24 *in* D. Capen, editors. The use of multivariate statistics in studies

of wildlife habitat. U.S. Forest Service, Rocky Mountain Forest and Range Experiment Station, General Technical Report RM-87, Fort Collins, Colorado.

Kerr, M. K., and G. A. Churchill. 2001. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. Proceedings of the National Academy of Sciences 98:8961–8965.

Kwak, T. J. and J. T. Patterson 2007. Community indices, parameters, and comparison. Pages 677–763 *in* C. S. Guy and M. L. Brown, editors. Analysis and interpretation of freshwater fisheries data. American Fisheries Society, Bethesda, Maryland.

Legendre, P., and L. Legendre. 1998. Numerical ecology, 2nd edition. Elsevier, Amsterdam.

Maechler, M., A. Struyf, M. Hubert, and K. Hornick. 2009. cluster: cluster analysis extended Rousseeuw et al. R library.

McKenna, Jr., J. E., B. M. Davis, M. C. Fabrizio, J. F. Savino, T. N. Todd, and M. Bur. 2008. Ichthyoplankton assemblages of coastal west-central Lake Erie and associated habitat characteristics. Journal of Great Lakes Research 34:755–769.

Mehner, T., K. Holmgren, T. L. Lauridsen, E. Jeppesen, and M. Kiekmann. 2007. Lake depth and geographical position modify lake fish assemblages of the European 'Central Plains' ecoregion. Freshwater Biology 52:2285–2297.

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.

Morris, C. C., T. P. Simon, and S. A. Newhouse. 2006. A local-scale in situ approach for stressor identification of biologically impaired aquatic systems. Environmental Contamination and Toxicology 50:325–334.

Nemec, A. F. L., and R. O. Brinkhurst. 1988. Using the bootstrap to assess statistical significance in the cluster analysis of species abundance data. Canadian Journal of Fisheries and Aquatic Sciences 45:965–970.

Oksanen, J., F. G. Blanchet, R. Kindt, P. Legendre, R. G. O'Hara, G. L. Simpson, P. Solymos, M. Henry, H. Stevens, and H. Wagner. 2009. vegan: Community ecology package. R library.

OMNR (Ontario Ministry of Natural Resources). 2007. Stream assessment protocol for southern Ontario. Ontario Ministry of Natural Resources, Picton.

R Development Core Team. 2010. R 2.10.1. A language and environment for statistical computing. R Roundation for Statistical Computing, Vienna, Austria.

Peres-Neto, P. R., D. A. Jackson, and K. M. Somers. 2003. Giving meaningful interpretation to ordination axes: assessing loading significance in principal component analysis. Ecology 84:2347–2363.

Peres-Neto, P. R., D. A. Jackson, and K. M. Somers. 2005. How many principal components? Stopping rules for determining the number of nontrivial axes revisited. Computational Statistics and Data Analysis 49:974–997.

Pillar, V. D. 1999. How sharp are classifications? Ecology 80:2508–2516.

Podani, J. 2000. Introduction to the exploration of multivariate biological data. Backhuys Publishers, Leiden, Netherlands.

Poos, M. S., S. C. Walker, and D. A. Jackson. 2009. Functional-diversity indices can be driven by methodological choices and species richness. Ecology 90:341–347.

Rousseeuw, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20:53–65.

Schaefer, J. A., and C. C. Wilson. 2002. The fuzzy structure of populations. Canadian Journal of Zoology 80:2235–2241.

Shimodaira, H. 2002. An approximately unbiased test of phylogenetic tree selection. Systematic Biology 51:492–508.

Sowa, S. P., G. Annis, M. E. Morey, and D. D. Diamond. 2007. A gap analysis and comprehensive conservation strategy for riverine ecosystems of Missouri. Ecological Monographs 77:301–334.

Suzuki, R. and H. Shimodaira. 2009. pvclust. Hierarchical clustering with *P*-values via multiscale bootstrap resampling. R library.

Tibshirani, R., G. Walther, and T. Hastie. 2001. Estimating the number of clusters in a dataset via the gap statistic. Journal of the Royal Statistical Society B 63:411–423.

Tilman, D. 1997. Distinguishing between the effects of species diversity and species composition. Oikos 80:185.

Tonidandel, S., and J. E. Overall. 2004. Determining the number of clusters by sampling with replacement. Psychological Methods 9:238–249.

Van Aelst, S., X. Wang, R. Zarnar, and R. Zhu. 2006. Linear grouping using orthogonal regression.

Computational Statistics and Data Analysis 50:1287–1312.

Winemiller, K. O., H. Lopez-Fernandez,, D. C. Taphorn, L. G. Nico, and A. Barbarino Duque. 2008. Fish assemblages of the Casiquiare River, a corridor and zoogeographical filter for dispersal between the Orinoco and Amazon basins. Journal of Biogeography 35:1551–1563.

Appendix A

Appendix 1. List of the species common and scientific names and the number of sites encountered in 20 sites from each the two watersheds sampled. Sydenham River denoted by S and Wilmot Creek by W.

Name		S	W	Name		S	W
American brook lamprey	*Lampetra appendix*	0	3	Black bullhead	*Ameirus melas*	7	0
Sea lamprey	*Petromyzon marinus*	0	2	Yellow bullhead	*A. natalis*	5	0
Longnose gar	*Lepisosteus osseus*	7	0	Brown bullhead	*A. nebulosus*	1	0
Gizzard shad	*Dorosoma cepedianum*		0	Channel catfish	*Ictalurus punctatus*	6	0
Coho salmon	*Oncorhynchus kisutch*	0	8	Stonecat	*Noturus flavus*	10	0
Rainbow trout	*O. mykiss*	0	20	Tadpole madtom	*N. gyrinus*	6	0
Chinook salmon	*O. tshawytscha*	0	6	Brindled madtom	*N. miurus*	3	0
Atlantic salmon	*Salmo salar*	0	9	Blackstripe topminnow	*Fundulus notatus*	3	1
Brown trout	*S. trutta*	0	18	Brook stickleback	*Culaea inconstans*	1	0
Brook trout	*Salvelinus fontinalis*	0	2	White perch	*Morone americana*	1	0
Northern pike	*Esox lucius*	5	0	White bass	*M. chrysops*	1	0
Chain pickerel	*E. niger*	0	1	Rock bass	*Ambloplites rupestris*	16	0
Goldfish	*Carassius auratus*	1	0	Green sunfish	*Lepomis cyanellus*	11	0
Spotfin shiner	*Cyprinella spiloptera*	15	0	Pumpkinseed	*L. gibbosus*	11	0
Common carp	*Cyprinus carpio*	8	0	Bluegill	*L. macrochirus*	2	0
Striped shiner	*Luxilus chrysocephalus*	1	0	Longear sunfish	*L. megalotis*	8	0
Common shiner	*L. cornutus*	9	0	Smallmouth bass	*Micropterus dolomieui*	4	0
Redfin shiner	*Lythrurus umbratilis*	11	0	Hornyhead chub	*Nocomis biguttatus*	1	0
Largemouth bass	*M. salmoides*	3	0	White crappie	*Pomoxis annularis*	6	0
Emerald shiner	*Notropis atherinoides*	3	0	Yellow perch	*Perca flavescens*	1	0
Ghost shiner	*N. buchanani*	3	0	Walleye	*Sander vitreus*	2	0
Spottail shiner	*N. hudsonius*	1	0	Eastern sand darter	*Ammocrypta pellucida*	3	0
Mimic shiner	*N. volucellus*	11	0	Greenside darter	*Etheostoma blennioides*	11	0
Northern redbelly dace	*Phoxinus eos*	4	0	Rainbow darter	*E. caeruleum*	0	6
Bluntnose minnow	*Pimephales notatus*	20	0	Fantail darter	*E. flabellare*	3	0
Fathead minnow	*P. promelas*	1	0	Least darter	*E. microperca*	6	0
Blacknose dace	*Rhinichthys atratus*	3	7	Johnny darter	*E. nigrum*	20	4
Longnose dace	*R. cataractae*	0	6	Logperch	*Percina caprodes*	7	0
Creek chub	*Semotilus atromaculatus*	9	5	Blackside darter	*P. maculata*	18	7
Quillback	*Carpiodes cyprinus*	1	0	Mottled sculpin	*Cottus bairdii*	0	17
white sucker	*Catostomus commersonii*	19	6	Slimy sculpin	*C. cognatus*	0	1
Northern hog sucker	*Hypentelium nigricans*	6	0				
Spotted sucker	*Minytrema melanops*	2	0				
Silver redhorse	*Moxostoma anisurum*	10	0				
golden redhorse	*M. erythrurum*	11	0				
Shorthead redhorse	*M. macrolepidotum*	9	0				
Greater redhorse	*M. valenciennesi*	2	0				