Using assemblage data in ecological indicators: A comparison and evaluation of commonly available statistical tools

Joseph M. Smith [a,*], Martha E. Mather [b,1]

[a] Massachusetts Cooperative Fish and Wildlife Research Unit, University of Massachusetts, Amherst, MA 01003, USA
[b] U. S. Geological Survey, Massachusetts Cooperative Fish and Wildlife Research Unit, University of Massachusetts, Amherst, MA 01003, USA

ARTICLE INFO

ABSTRACT

Ecological indicators are science-based tools used to assess how human activities have impacted environmental resources. For monitoring and environmental assessment, existing species assemblage data can be used to make these comparisons through time or across sites. An impediment to using assemblage data, however, is that these data are complex and need to be simplified in an ecologically meaningful way. Because multivariate statistics are mathematical relationships, statistical groupings may not make ecological sense and will not have utility as indicators. Our goal was to define a process to select defensible and ecologically interpretable statistical simplifications of assemblage data in which researchers and managers can have confidence. For this, we chose a suite of statistical methods, compared the groupings that resulted from these analyses, identified convergence among groupings, then we interpreted the groupings using species and ecological guilds. When we tested this approach using a statewide stream fish dataset, not all statistical methods worked equally well. For our dataset, logistic regression (Log), detrended correspondence analysis (DCA), cluster analysis (CL), and non-metric multidimensional scaling (NMDS) provided consistent, simplified output. Specifically, the Log, DCA, CL-1, and NMDS-1 groupings were ≥60% similar to each other, overlapped with the fluvial-specialist ecological guild, and contained a common subset of species. Groupings based on number of species (e.g., Log, DCA, CL and NMDS) outperformed groupings based on abundance [e.g., principal components analysis (PCA) and Poisson regression]. Although the specific methods that worked on our test dataset have generality, here we are advocating a process (e.g., identifying convergent groupings with redundant species composition that are ecologically interpretable) rather than the automatic use of any single statistical tool. We summarize this process in step-by-step guidance for the future use of these commonly available ecological and statistical methods in preparing assemblage data for use in ecological indicators.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

Scientists and managers need science-based methods to assess how human activities will impact resources. For example, environmental professionals are often asked to evaluate impacts of specific human activities on aquatic biota (Filipe et al., 2002; Anderson et al., 2006; Arthington et al., 2006). Ecological indicators that reflect the composition, structure, and function of ecosystems (Reza and Abdullah, 2011) can meet this need for science-based methods that aid impact assessment. Biological assemblage datasets are widely available and can potentially provide useful information on background ecosystem conditions and ecological responses to anthropogenic impacts. For example, the Environmental Protec-

tion Agency's guidelines for ecological indicator suitability include conceptual relevance, feasibility of implementation, response variability, interpretation, and utility (Kurtz et al., 2001). Several impediments exist, however, to the use of assemblage datasets by environmental professionals. These complex data are composed of tens of species and hundreds of individuals so these data need to be simplified before they can be used in ecological indicator development and testing. In addition, the ideal data processing protocol should reduce the complexity of the assemblage data, yet still retain inherent ecological information. Finally, the output from statistical methods commonly used to process assemblage data are rarely compared and evaluated, so little consensus exists on which approach works best. Consequently, field researchers and managers who construct and test ecological indicators could benefit from a practical guide to using assemblage data that includes a systematic comparison of commonly used statistical methods coupled with an evaluation of whether the resulting simplified statistical groupings reflect ecological patterns in the original dataset.

* Corresponding author. Tel.: +1 413 230 0244; fax: +1 413 545 4358.
E-mail addresses: jsmith@eco.umass.edu (J.M. Smith), mather@eco.umass.edu (M.E. Mather).
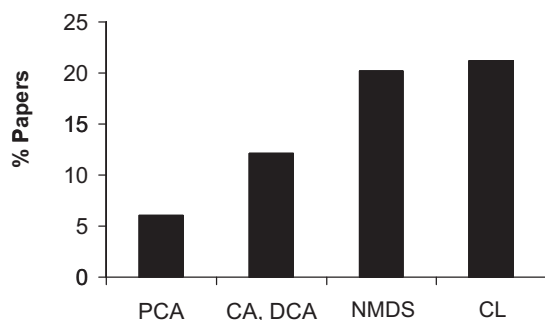[1] Tel.: +1 413 329 2105; fax: +1 413 545 4358.

Fig. 1. Percentage of papers that used common statistical grouping methods to summarize fish assemblages in a literature review of studies (*N* = 101) from 2005 to 2010. The review was conducted using Web of Science. The terms "fish community(ies)" or "fish assemblage(s)" were searched in the journal article title, and the words "stream" or "river" within the topic field. Only papers that actively sampled fish in freshwater streams were included in the review. PCA = principal components analysis, CA = correspondence analysis, DCA = detrended correspondence analysis, NMDS = non-metric multidimensional scaling, and CL = cluster analysis. Papers are listed in Supplementary Material – Literature Table. These four analyses collectively comprise almost 60% of the analytical tools used in this database.

In the past, investigators have used a variety of statistical or ecological methods to process assemblage data. Commonly used multivariate statistical analyses that group species data include: cluster analysis (CL) (e.g., Orrego et al., 2009; Penczak et al., 2009), non-metric multidimensional scaling (NMDS) (e.g., Chick et al., 2006; Lorion and Kennedy, 2009), correspondence and detrended correspondence analyses (CA, DCA) (e.g., Falke and Gido, 2006; Humpl and Pivnička, 2006), and principal components analysis (PCA) (e.g., Lamouroux and Cattaneo, 2006; Orrego et al., 2009) (Fig. 1; Supplementary Material – Literature Table). Because multivariate statistics transform biological data into mathematical relationships, an effective processing protocol needs to include some validation that the simplified assemblage data reflect ecological trends. Guilds are a common ecological-based approach to grouping species (Welcomme et al., 2006; Noble et al., 2007). Concurrently using guild classifications, species information and quantitative approaches can provide insights into whether the simplified groupings that result from the multivariate and other grouping statistics reflect known ecological relationships.

Many researchers use exploratory analyses from biotic community datasets to develop ecological indicators (Podani and Csányi, 2010). How assemblage data are used in ecological indicators may influence the most appropriate statistical analysis. For example, assemblage data can be simplified to produce fewer, ecologically meaningful multispecies groupings. In addition, these simplified assemblage groupings can be used to relate biota to environmental stressors of interest to managers or to identify influential species. Fish are often used as ecological indicators because of their links to environmental conditions (Kanno et al., 2009; Maceda-Veiga and De Sostoa, 2011). Here we used a statewide stream database as a case study to illustrate how common statistical methods can be used for these purposes. Specifically, we sought to provide a standardized framework for transforming raw data into useable ecological metrics by addressing the following objectives. First, we compared six common statistical grouping methods [i.e., cluster analysis (CL), non-metric multidimensional scaling (NMDS), principal components analysis (PCA), detrended correspondence analysis (DCA), logistic regression (Log), Poisson regression (Poisson)]. Although a number of novel statistical tools are currently being developed, we chose to compare well-developed statistical methods because these are readily available and most likely to be used by environmental professionals. We then evaluated the groupings that resulted from these statistical methods by comparing convergence, redundancy, and statistical similarity among

methods, using guilds to interpret these groupings ecologically, and examining individual species to provide ecological validation. Second, we evaluated if the biotic groupings from these commonly used statistical methods could be related to environmental stressors. For this, we used low streamflow as an example of a stressor of interest in environmental management. However, any environmental variable of interest to regulators could be used. Third, we examined the distribution of individual species within each grouping across methods to identify influential taxa. Finally, we recommend a practical, step-by-step process that researchers and managers can use with other assemblage datasets. Many excellent statistical treatments of multivariate analyses already exist in the literature; here, rather than duplicate these efforts, we seek to provide a user-friendly approach for field-oriented environmental professionals who want to use assemblage data.

2. Materials and methods

2.1. Dataset

Fish were sampled at 344 sites throughout Massachusetts (Fig. 2A). The sites were located within two U.S. Environmental Protection Agency (EPA) level III ecoregions, the Northeastern Highlands (58) and the Northeastern Coastal Zone (59; Fig. 2A). Sites were small, wadeable streams with drainage areas from 0.04 to 262.93 km^2, mostly ≤10 km^2 (Fig. 2B). All fish species sampled were present in both ecoregions. In this Massachusetts Division of Fisheries and Wildlife dataset, fish were sampled from June to August 1998 to 2005. All sample reaches were at least 30 times the width of the stream, a length within the range recommended for stream fish sampling (Simonson and Lyons, 1995). A single pass was made at all sites with one backpack electroshocker, a level of effort that has been shown to adequately characterize fish communities (Reid et al., 2009). Fish were identified to species, counted, measured, and returned to the stream. Species that were present in less than 5% of all samples were removed from the analyses to limit the undue influence of rare species (Gauch, 1982).

2.2. Create groupings that simplify assemblage data

We compared six statistical methods of which three used assemblage data only and three used assemblage plus environmental data (Table 1). As a result of each of these statistical methods, one or more distinct "groupings" of multiple species were identified. Throughout, "groupings" refers to the species aggregations that result within and across multivariate analyses (e.g., CL-1, CL-2, NMDS-1, NMDS-2; Table 1). Most of these statistical methods can be calculated using the number of individuals (N) or the number of species (S). The first three methods that used assemblage data only included: cluster analysis (CL), non-metric multidimensional scaling (NMDS), and principal components analysis (PCA). We also chose three additional statistical tools that combined assemblage and environmental data: detrended correspondence analysis (DCA), logistic regression (Log), and Poisson regression (Poi).

Cluster analysis is a classification technique that places objects that are sufficiently similar into the same group while identifying distinctions between groupings (Legendre and Legendre, 1998). We used hierarchical agglomerative clustering with a Bray-Curtis resemblance matrix based on species presence-absence at each site (PRIMER 6.1.10 software; Clarke and Gorley, 2006). For this, we used a group average clustering algorithm (Legendre and Legendre, 1998). Species that were at least 30% similar were included in the cluster groups (CL-1, CL-2) and were calculated using species (S) and abundance (N) data (Table 1).