Addressing the removal of rare species in multivariate bioassessments: The impact of methodological choices

Mark S. Poos*, Donald A. Jackson

Department of Ecology and Evolutionary Biology, University of Toronto, 25 Harbord Street, Toronto, Ontario, Canada M5S 3G5

ARTICLE INFO

ABSTRACT

Multivariate analyses are important tools for the biological assessment of ecological communities. Despite the popularity of multivariate analyses in bioassessments, there is considerable controversy over how to treat rare species. As this debate remains unresolved, the objective of this study was to develop a methodology to quantify the impacts of removing rare species relative to other decisions inherent in multivariate analyses and to provide insight into their relative influence in our studies. Using fish species from a well-sampled system, we assessed the impact of several choices common to multivariate analyses, including the removal of rare species, ordination technique and measures of multivariate resemblance. Comparisons of multivariate analyses demonstrated the choice of ordination method explained 26% of the variation among the various results, followed by the choices regarding the removal of rare species (24.8%) and resemblance measure (11%). At the same time, the removal of rare species had important site-level impacts relative to full dataset, including a >9 fold change in sites impacted by the removal of single species, with an emphasis on removing species more correlated to anthropogenic stress. Our study demonstrates that the removal of rare species had similar or greater influence in multivariate analyses as other choices inherent in their calculation, such as the choice of ordination method. Better justifications for the removal of rare species, along with all decisions in multivariate analyses, are needed to move bioassessments forward.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

Multivariate analyses have become important tools in biological assessments of aquatic communities (Norris, 1995; Wright et al., 2000). Multivariate analyses are widely used and have been shown to be accurate and sensitive ways to quantify biological impacts and recoveries (Linke et al., 2005). Several national bioassessment programs are based on multivariate measures, including those in the United Kingdom (e.g., RIPVACS; Wright et al., 2000) Australia (AUSRIVAS; Simpson and Norris, 2000), Canada (Reynoldson et al., 2001), as well as elsewhere (Bailey et al., 2004; Van Sickle et al., 2007). In total, approximately 30% of bioassessment articles use multivariate statistical methods (Dolédec and Stazner, 2010).

The application of multivariate analyses to bioassessments of aquatic communities has been a major point of controversy. Polarized debates remain regarding the use of multimetric versus multivariate methods (Dolédec and Stazner, 2010), the taxonomic resolution needed for species identification (Arscott et al., 2006), and to some extent, the use of specific analytical methods (Marchant et al., 2006; Van Sickle et al., 2007). In particular, the treatment of rare species has received much attention in bioassessment studies (e.g., Faith and Norris, 1989; Norris, 1995; Cao et al., 1998; Marchant, 1999, 2002). On one hand, researchers often remove rare species with the perceived notion that they may add noise to multivariate analyses and provide little additional information beyond more common species (Gauch, 1982; McCune and Grace, 2002). On the other hand, researchers often retain rare species in multivariate analyses because they may be better indicators of ecosystem stress than common species (Faith and Norris, 1989; Cao et al., 1998, 2001), given the assumption that some or all of these rare species may be more sensitive to the stressor(s). In either case, the debate regarding the treatment of rare species has remained unresolved and researchers need to be aware of the impact of their decision of how to treat rare species (among others).

There are many difficulties in attempting to resolve the debate regarding the treatment of rare species in bioassessments and in other types of multivariate studies. For example, most multivariate approaches require several more decisions beyond whether or not to remove rare species, and these may reduce insight into the effect of rare species on resultant analyses. Researchers using multivariate methods typically must choose a type of resemblance measure (i.e., a distance or similarity coefficient) and ordination technique,

* Corresponding author. Current address: Fisheries and Oceans Canada, Great Lakes Laboratory for Fisheries and Aquatic Sciences, 867 Lakeshore Road, Burlington, Ontario, Canada L7R 4A6. Tel.: +1 905 319 7249; fax: +1 905 336 6437.
   E-mail addresses: Mark.Poos@utoronto.ca, mark.poos@dfo-mpo.gc.ca (M.S. Poos).

where such choices have been shown to significantly alter results (Podani, 2000; Hirst and Jackson, 2007; Poos et al., 2009). As such, the resolution to the debate regarding the impact of rare species cannot proceed until the effect of removing rare species is placed into a context comparable to other decisions inherent in multivariate bioassessments. Unfortunately relatively little effort has gone into comparing these methods of bioassessments (Norris, 1995; Marchant et al., 2006) and few studies have viewed decisions in analyses in a holistic manner (e.g., how do all decisions inherent in multivariate bioassessments impact results and which ones are of greater importance?).

There are two general arguments for the inclusion or exclusion of rare species in multivariate bioassessments. The first argument for removing rare species from bioassessments is that rare species provide limited interpretive value (Marchant, 1999). Proponents of this argument suggest that rare species may simply reflect stochastic sampling effects and therefore add noise rather than information, to the statistical solution (Gauch, 1982; Clarke and Green, 1988; Bailey et al., 2004). We refer to this argument as the statistical impact argument. Support for this argument has come from work in previous decades, which showed results from multivariate methods could be driven by the inclusion of rare species alone (see Cao et al., 2001; Dolédec and Stazner, 2010). To some degree this argument has been examined in the literature with analyses of certain aspects of developing a multivariate bioassessment, like data standardizations (Jackson, 1993a; Cao et al., 1999), resemblance coefficients (Jackson et al., 1989), ordination method (Marchant, 1990), or their combinations (e.g., data standardization and resemblance coefficients; Jackson, 1993a; Hirst and Jackson, 2007; taxonomic resolution and rarity; Arscott et al., 2006). Unfortunately, what has remained largely absent is a quantitative evaluation of the role of rare species in community assessments, including an evaluation which answers the practical question of how important rare species are relative to other decisions in multivariate analyses (but see Faith and Norris, 1989). In this context, the statistical argument can be tested as a hypothesis, with the prediction that differences among analyses should show considerable variation where rare species are removed relative to analyses where other decisions have been altered (e.g., resemblance coefficient, ordination method). Specifically, does the inclusion or exclusion of rare species lead to greater changes in community analyses than those arising due to other decisions (e.g., the well documented differences due to choice of ordination)?

The second argument for the inclusion or exclusion of rare species is that such species should be included in multivariate analyses because they are better indicators of ecosystem stress than are common species (Cao et al., 1999, 2001), i.e., common species tend to have broad ranges of tolerance to many conditions and therefore may not provide as good indicators. We refer to this argument as the biological impact argument. Support for this argument has come from empirical studies which note the importance of including rare species for conservation issues (Cao et al., 1999, 2001). Proponents of this argument suggest that exclusion of rare species may lead to an underestimation of differences between impacted and un-impacted sites (Cao et al., 1999, 2001). As one of the main goals of a bioassessment is to determine site-level impacts (Barbour et al., 1999; Wright et al., 2000), this argument assumes that the loss of rare species represent the loss of the strongest signals of impact, via decreases in species diversity or changes in community composition (Cao et al., 1998). Therefore the biological impact argument can also be tested as a hypothesis, with the prediction that sites where rare species were removed should be more affected across multivariate analyses than sites that had no species removed. If this result is not found, it would indicate that the exclusion of rare species is warranted as they do not provide meaningful information beyond that captured by more common species.

Given these definitions and hypotheses, the objective of our study was to determine the biological and statistical impact of removing rare species relative to other methodological decisions inherent in multivariate analyses (e.g., choice of ordination method and resemblance coefficient). For this evaluation, we use data of fish communities from of the Sydenham River as a model system because it has the highest diversity of aquatic fauna in Canada, as well as the highest number of species at risk in Canada (Poos et al., 2010). The Sydenham River is a 2725 km^2 watershed, which drains into Lake St. Clair, a connecting waterway of the Laurentian Great Lakes, between Lake Huron and Lake Erie (Poos et al., 2007). Although the human population is small (∼100,000 within the watershed), anthropogenic impacts of turbidity remain an important concern, mostly due to land-use activities dominated by agriculture (Poos et al., 2010). The Sydenham River has undergone detailed sampling (Poos et al., 2007, 2008) which provides a high-quality dataset. As there are many types of rarity (Gaston, 1994), we define rare species as those which occur infrequently (i.e., at few locations or low prevalence). As well, we define species that have a conservation designation (e.g., endangered, threatened, special concern), as species at risk.

2. Materials and methods

Fishes were collected from the Sydenham River using the Ontario Stream Assessment Protocol (OMNR, 2007). The Ontario Stream Assessment Protocol represents a typical bioassessment protocol for monitoring impacts to aquatic systems (e.g., Barbour et al., 1999; Wright et al., 2000) and has been used effectively to monitor changes in riverine communities (Poos et al., 2008). We used fish rather than benthic macroinvertebrates as they are relatively easy to identify and enumerate and also have been used extensively in multivariate bioassessments (Bailey et al., 2004; Dolédec and Stazner, 2010); however the approach we propose is equally suited to macroinvertebrates or any other taxonomic group of species.

Fishes were sampled using a variety of approaches (see Poos et al., 2007); however, for this analysis only electrofishing data were used as it is the most commonly used method for the biomonitoring protocols and regarded as the most effective gear type for sampling stream-fish assemblages and fish species at risk (Bohlin et al., 1989; Reynolds, 1996; Poos et al., 2007). Sampling sites were chosen at random across the entire Sydenham watershed. Species were collected at 50 sites in 2002 and 25 additional sites in 2003.

2.1. Evaluating decisions in multivariate bioassessments

Prior to analysis, four treatments of the removal of species were applied to the site-by-species matrix. Traditionally, researchers arbitrarily decide which characteristics define a rare species within a sample (Faith and Norris, 1989; McCune and Grace, 2002). Some researchers suggest eliminating species which occur at single sites because of the inflated correlations created by attempting to relate potentially random features at that site to its lone occurrence (Legendre and Legendre, 1998). Others suggest removing species that occur at less than five percent (McGarigal et al., 2000) or ten percent of sites (Marchant, 1990; McCune and Grace, 2002) or at even higher thresholds (Marchant et al., 1997). We used the treatments of: (1) the full dataset (All), (2) removing single-occurrence species (M1); (3) removing species found at less than five percent of sites (M5); and, (4) removing species found in less than 10% of sites (M10). These criteria represented 0, 4, 9, and 21 species, respectively of the 67 species dataset.

Resemblance coefficients were calculated from each of the four site-by-species matrices of which rarely sampled species were