Faut-il exclure ou non les espèces rares d'une matrice de données avant une analyse des correspondances?

Les textes ci-dessous sont des extraits d'une discussion qui a eu lieu sur le groupe ORDNEWS (ordnews@colostate.edu) en octobre 2002. Le premier message est la question posée, et les suivants sont les contributions de plusieurs abonnés à la liste. On constatera que les solutions sont multiples, et dépendent des questions posées, de la nature des données, du but de l'analyse... et des opinions!

Daniel Borcard

I plan to perform a CCA on a large dataset containing 2,603,961 fish collected throughout an estuary. At the moment, I am trying to narrow down the 150 separate species in this dataset for inclusion in the analysis to reduce the effect of rare species. My dataset is pretty extensive (3283 samples).

It looks like many researchers narrow down rare species by using only those that represent 1% or more of the total number of individuals in the dataset. If I only use species representing above 1%, my data is narrowed down to 7 species. Almost all of these species are schooling fish that are ubiquitous within the estuary (one species represented 57% of all individuals in dataset).

Shannon D. Whaley, Florida Marine Research Institute, St Petersburg, Florida, USA

===

First of all, I would try a strong transformation on the data set to see whether by reducing the scale of relative variation among your variables (species) you do override numerical over-representation from schooling fishes.

Second, in species reductions, a way of avoinding the loss of a vast amount of "rare" species (i.e. of information) is to remove those species not reaching a given percentage of individuals PER SAMPLING UNIT rather than considering the whole data set.

If that doesn't work nicely, I would very much consider the option of analysing CCA patterns on schooling fish species (7 species?; ordination 1), on one side, and on the rest of species (143 species?; ordination 2), on the other side.

Salvador Herrando-Pérez, Castellón de la Plana, Espagne

Unless there is a true limitation of computer or program capacity I would always include all of the rare species. Although rare species are often fortuitous and not a controlling factor, they can be both indicators, and in my experience sometimes a suite of rare species forms a group that does become a controlling or indicating factor. In short, run it with and without and ask yourself what the difference in the outcomes is. If there is no difference then the significance of the rare ecies becomes moot.

Patrick Murphy
Plant Ecologist

I have had similar problems with some of our datasets here. In Lake Texoma, >70% of the individuals from our shoreline seining samples over 3 years have consisted of inland silversides Menidia beryllina and threadfin shad Dorosoma petenense (total of 48 spp were collected). A situation like this make everything rare. Rather than eliminating rare species based solely on %abundance, maybe consider incidence of occurrence.

Michael Eggleton

Exclusion of 'rare species' can be questioned for several reasons.

1) Species-abundance distributions or patch-occupancy distributions often differ among different sites/assemblages, some naturally having more rare species than others. A cut-off line, e.g., 0.5% frequency or 0.5% of site total abundance, necessarily affect different sites differentially. For example, it may remove 5% of all species recorded at Site A, but up to 50% at Site B . . .
2) If the goal of ordination is to recognize a major environmental gradient or to classify assemblages into major types, inclusion/exclusion of rare species is likely to make little difference. However, if one is trying to detect/assess environmental changes and human impacts, exclusion of rare species could be a problem (see Cao et al. 2001. JNABS 21:144-153).
3) When we talk about 'rare species', it should be kept in mind that we are talking about 50-80% of the local or regional species pool, depending on the cut-off lines,. . . . . we need to consider this issue not only from the standpoint of statistics, but also from ecological/biological aspects.
4) Rarity takes a variety of forms, e.g, low abundance vs. low patch-occupancy . . . ., and many factors influence the rarity observed, such as sampling effort, sampling times, . . .

Yong Cao, Utah State University, USA

Just another aspect to consider, particularly when working with fish data: sampling seletivity. Many species in the sample will be rare not because they are rare in the studied area, but because the collecting method is not efficient for capturing them. Therefore, they do not represent true rarity patterns and, depending on your objectives, could lead to bias in interpreting the results.
If excluding species from the analysis is really an option, maybe those species should be first candidates.

Fernando G. Becker, Fundação Zoobotânica do Rio Grade do Sul Brazil

----

I take a technical point of view here: Why rare species are removed in *Correspondence* *Analysis* (CA) -- including CCA as a special case. If we take this point of view, this idea is not very good: CA is said to be sensitive to rare species, but that means species that occur in only one or two sampling units (points, plots, locations what ever). The total number of individuals is not so important.
Another point is that it looks to me that CA is much less sensitive to rare species than people often think. It is true that rare species come out extreme in ordinations, but they do not influence the ordination of sites so much. CA is a weighted ordination of Chi-square dissimilarities (not `kind of' but exactly). Rare species do have high Chi-square distances from the origin, but they have low weights. A peculiarity in CA is that site ordination may change little though species ordinations vary wildly. You got inspect both separately.

However, when you are looking at site profiles, the very abundant schooling fish will dominate -- unless they happen to be absent just in that point. So your rare species will probably have only a minute effect in site profiles. If you want to boost their influence in ordination, you should consider some transformation (sqrt, a higher root)...

Jari Oksanen, Oulu, Finlande

The question of species removal is a part of the more general question of how to balance quantitative and qualitative aspects of species performance.(...) In ordination, the species data is used to place samples along underlying gradients. Thus, the weight attributed to a species via its abundance should reflect the species' value as an indicator of gradient position for the sample in which the species occurs. Thus:
1) Removal of rare species accords with the view that rare species are without value as indicators of the gradient position of the samples in which they occur.
2) Using the raw individual counts as input to ordination accords with the view that a species' value as indicator of gradient position increases with number of individuals.
I suppose you disagree with both of these statements. The solution to this problem chosen by most plant ecologists is (a) not to remove rare species because that all convey some information of ecological interest, and (b) to transform (weigh) the raw data so that a balance between the quantitative (abundance) and the qualitative (presence) information is achieved. This is done by weighing, i.e. adjustment of the scale for recorded abundance. There are several ways to do this, e.g. by the power function (see van der Maarel 1979, in Vegetatio 39: 97-114; and R. Økland 1990, in Sommerfeltia Suppl. 1: 1-233). A range of the abundance scale (ratio of maximum recorded abundance vs just presence) of 5-10 is often recommended as an ecologically sound compromise.

Rune H. Økland, Oslo, Norvège

Another way of considering rare species in your analysis is based on a functional approach, that is, looking for what species do rather than what they are.
I have found the functional approach very usefull in two case studies focused on marine macrobenthos, where I looked at 1/ feeding guilds in assessing the impact of fly-ash (a non-organic waste) sublittoral dumping in the North Sea (Sarsia 2001, 389-400), and 2/ reproductive guilds in evaluating occurrence and dispersion patterns of island invertebrates in the Canary Islands (Cahiers de Biologie Marine 2001, 275-287).
Here, species are allocated to functional groups and abundant values correspond with those of the multi-species functional groups. This obvisously 1/ relies on the information available in the literature for species life history traits, and 2/ may go hand in hand with your research goals since each study variable may account for a distinct ecological process. I guess that schooling fishes have very particular life history features, and that some of your rare species may become functionally relevant in assessing overall patterns of community functioning.

Salvador Herrando-Pérez, Castellón de la Plana, Espagne

There is evidence that the exclusion of rare species in ordinations can affect the recovery of key underlying environmental gradients; see for example:
Faith DP, Norris RH. 1989. Correlation of environmental variables with patterns of distribution and abundance of common and rare fresh-water macroinvertebrates. Biol. Conserv. 50 (1-4): 77-98.

Dan Faith
