Vol. 143: 303–306, 1996 | MARINE ECOLOGY PROGRESS SERIES Mar Ecol Prog Ser | Published November 14

NOTE

Assessing the relationship between multivariate community structure and environmental variables

J. A. F. Diniz-Filho[1], L. M. Bini[2]

[1]Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Goiás. CP 131, 74.001-970, Goiânia, Goiás, Brazil
[2]Programa de Pós-Graduação em Ciências da Engenharia Ambiental, USP, and Nupelia, Universidade Estadual de Maringá, 87.020-900, Maringá, Paraná, Brazil

ABSTRACT: Clarke & Ainsworth's (1993; Mar Ecol Prog Ser 92:205–219) method of linking multivariate community structure to environmental variables possesses an heuristic and conceptual interest. However, we believe that there are other strategies which require less computational effort and also permit an accurate statistical test of the relationship between multivariate community structure and environmental variables. The objective of this communication is to discuss some of these many strategies, emphasizing the flexibility of Mantel test design for this task.

KEY WORDS: Assessing community structure-environment relationship

Multivariate analysis has been widely used in the last 30 yr in ecology and systematics to solve a very distinct array of problems (Legendre & Legendre 1983, Krebs 1989, James & McCulloch 1990). Among these, the most common approaches involve (1) the analysis of association between (pairs of) species, reflecting niche or trophic patterns and ecological interactions such as competition, mutualism and predation, and (2) the analysis of similarity between samples, based on overall community analysed, which must be explained through distribution of environmental variables or intrinsic factors, such as larger migration rates between neighbouring samples.

There are many possible strategies for evaluating the relationship between multivariate community structure and environmental variables, and the choice of one of these relies on the presence of some analytical (mathematical) constraints in data, available computer software and, of course, knowledge about them.

In a recent paper, Clarke & Ainsworth (1993) proposed a new method for investigating this relationship, based on matrix comparison and multidimensional scaling. They expected to answer the following 2 questions: (1) How well is the community structure explained by the full set of environmental variables measured? and (2) Which variables are redundant in the sense of failing to strengthen the 'explanation' of biotic patterns once certain other variables are taken into account? The method they proposed consists of the following: (1) biotic and abiotic data matrices are handled separately, initially transforming each according to the needs of the differing similarity measures; (2) the among-sample similarity matrix for the biota is constructed only once but the equivalent triangular matrix for the abiotic data is computed many times, in fact for all possible combinations of environmental variables at each 'level of complexity' of explanation (variables taken singly, 2 at a time, 3 at a time, etc.); (3) the rank correlation (e.g. ρ_w, weighted Spearman coefficient) between the biotic and abiotic triangular matrices is calculated in every case. The highest few coefficients at each level of complexity are tabulated, allowing the extent of improvement or deterioration in the match to be traced as further variables are added; and (4) the final step is to display the biotic MDS (non-metric multidimensional scaling) in conjunction with ordinations of the most important environmental variable combinations.

Although Clarke & Ainsworth's (1993) method possesses heuristic and conceptual interest, we believe that there are other strategies which require less computational effort and also permit an accurate statistical test of the relationship. The objective of this communication is to discuss some of these many strategies, emphasizing the flexibility of Mantel test design to solve these problems.

Basic concepts. First, it is necessary to understand the basic nature of both multivariate methods and data in ecology. Multivariate data consist of a matrix of p variables measured in n samples. These p variables can be partitioned into 'community' variables (relative

abundances of species) and environmental variables, which can also be subsequently subdivided into other groups, as will be discussed later On the other hand, multivariate methods include a large complex of techniques, used for distinct purposes. It is important to note, however, that methods based on similarity between pairs of samples (or species; Q- and R-mode analyses), including both ordination and clustering, are used only to permit the visualization of the relationships that are mathematically defined in a p-dimensional space, in a low dimensional space (usually no more than 3 axes). This dimensional reduction, of course, involves some distortions in the relationships between samples, which can be measured by stress and/or co-phenetic or matrix correlation between original similarity and similarity in the reduced space of dendrograms and ordinations (Sneath & Sokal 1973). As a logical consequence, evaluation of the relationship between multivariate community structure and environmental variables, based on the original similarity matrix, is by definition better than evaluations based on clustering and ordination results, such as Procrustes methods (Gower 1971) or estimating consensus trees between dendrograms (Rohlf 1974, 1982), both of which could be used to compare outputs of 2 multivariate analyses of the same samples (using community and environmental variables).

There are some multivariate methods that are not based on similarity matrices and are specifically designed to assess the relationship between 2 sets of data. The well known Canonical Correlation Analysis and its derivative, the Redundancy Analysis, are designed to find linear combinations in the 2 data sets, in such a way that the Pearson product-moment correlation between derived canonical scores reaches the maximum (Harris 1975, Johnson & Wichern 1992). Although largely applied to ecology and systematics, they have some limitations related to assumptions of linearity and multivariate normal distribution in data. More importantly, there must be a large number of observations in relation to the number of variables in the 2 sets, which is usually difficult in community analysis, when a large number of species (variables) is collected. Multivariate normal distribution and linearity may be overcome by using Canonical Correspondence Analysis (ter Braak 1986, 1987) which in turn assumes that a species' abundance is a unimodal function of position along environmental gradients (ter Braak & Prentice 1988, Palmer 1993). Both of these techniques are in some situations very difficult to interpret due to their mathematical constraints (Legendre & Legendre 1983). The more serious problem with a small number of observations in relation to variables available can be eliminated by using the recently developed Co-inertia Analysis (Dolédec & Chessel 1994), but this technique is too new to provide a definitive solution to this problem.

The most commonly used evaluation of multivariate community structure in ecology, using a single data matrix of relative abundances, starts with constructing similarity or dissimilarity matrices, using many algorithms, such as Morisita-Horn, Bray-Curtis, Canberra, etc. (Lamont & Grant 1979, Wolda 1981, Legendre & Legendre 1983, Washington 1984, Gower & Legendre 1986). The next step is usually to perform hierarchical clustering, such as UPGMA (Sneath & Sokal 1973), or ordination techniques, such as Principal Coordinate (PCOORD) (Gower 1966, Sneath & Sokal 1973) or Non-Metric Multidimensional Scaling (NMDS) (Kruskal 1964a, b), on the similarity matrix. A similar solution is to perform the well known Principal Component Analysis (PCA) directly on the data matrix, although scores will be distributed under a Euclidean metric in dimensionally reduced space. In all cases, the purpose is to observe relationships in a lower number of dimensions, since it is impossible to evaluate the original similarity matrix, which is defined in a p-dimensional space. Criticisms of these linear techniques have been made, and non-linear alternatives, such as detrended correspondence analysis, have been suggested as better solutions for ecological ordination, even better than NMDS (Gauch 1987). These criticisms, and specially the advantage of detrended correspondence analysis in relation to NMDS, however, cannot be considered concrete resolutions (Wartenberg et al. 1987, Peet et al. 1988, Palmer 1993). An important next step is to interpret the distribution of samples as a function of other ecological factors, usually the spatial arrangements of samples (in a spatial context, a useful null hypothesis is that similarity is proportional to geographic distance among samples; Sokal & Wartenberg 1981) or the correspondence with the distribution of some environmental variables. This step, however, is usually performed in a subjective way. When there are some environmental variables measured, a common strategy is to apply the canonical analyses already discussed, or simply to use principal component/coordinate scores as a response variable in a multiple regression in which environmental variables are used as predictor variables (Sokal & Unnasch 1988).

A general approach. We believe, however, that the best strategy to investigate the relationship between multivariate community structure and environmental variables is to apply the Mantel test (Mantel 1967, Mantel & Valand 1970), under many possible designs. The Mantel statistic (Z) for matrix correspondence is given as:

$$Z = \sum_i \sum_j E_{ij} M_{ij}$$

where E_{ij} and M_{ij} are the ith and jth elements of the symmetric matrices **M** and **E** to be compared, i.e. a

matrix of Morisita-Horn similarity between samples (**M**), based on relative abundances of p species and a Euclidean distance (**E**) between the same samples but using environmenal variables. Although Mantel (1967) and Mantel & Valand (1970) tested the statistical significance of *Z* based on a normal distribution of coefficients, this approach is biased (Mielke 1978). The most common approach today is to test its significance by randomization (Sokal 1979, Dietz 1983, Manly 1986a, b, 1991, Crowley 1992), in which rows and lines of one the matrices (e.g. **E**) are randomly permuted many times, and for each of them a new *Z* is estimated. After a large number of randomizations, the observed *Z* is then compared with an empirical distribution under the null hypothesis of no relationship between matrices. The number of permutations must be chosen according to the number of rows in the similarity matrix and the significance level desired for testing, and most applications use around 1000 permutations. However, a recent study shows that stability is only achieved with a much larger number, no less than 10 000 (up to 100 000) (Jackson & Somers 1989). In computational terms, the Mantel test is available in the NTSYS-PC ('Numerical Taxonomy and Multivariate Analysis System') package (Rohlf 1989), in 'Randomization Test' programs (Manly 1991) and in the R-Package (Legendre & Fortin 1989).

The *Z*-statistic, however, is not a coefficient, and only tells if matrices are associated or not. The magnitude of *Z* is highly dependent on the number of elements in the matrix and on their magnitudes (Smouse et al. 1986). However, it is easy to see that *Z* has a monotonic relationship (simultaneous increase or decrease) with Pearson correlation between matrices (matrix correlation), because in fact if **M** and **E** are standardized prior to the analysis, *Z* is the covariance between them (the numerator of Pearson correlation). Since all other elements in the formula of Pearson correlation (variances and means of **M** and **E**) are invariant with permutations, the monotonic relationship permits testing of the significance of Pearson correlation between matrices using *Z*-statistics. So, correlation between matrices, in which a Pearson's r is calculated, can be understood as a standardized *Z*-statistic (Smouse et al. 1986). Extending the argument, an obvious conclusion is that r^2 gives the proportion of **M** which is explained by **E**, analogous to the first question proposed by Clarke & Ainsworth (1993), which is then easily answered.

The Mantel test for matrix correspondence can also be easily extended to more than 2 matrices, using multiple regression (Manly 1986b), partial correlation (Smouse et al. 1986b, Fromentin et al. 1993) and path analysis (Taylor & Gottelli 1994) approaches (see Oden & Sokal 1992 for a recent review of multiple matrix Mantel tests). This way, the second question proposed by Clarke & Ainsworth (1993) can be answered, for example, with a multiple regression design, in which partial regression coefficients and the squared multiple correlation coefficient (R^2) are estimated using community matrix as a response variable and groups of environmental variables as predictor variable in a matrix randomization design. The remaining question is how to perform the partition of environmental variables to be used to define the predictor matrices. We suggest at least 4 alternatives: (1) divide the environmental variables into 'natural' groups, such as geographic variables (latitude, longitude), macroclimatic data, etc.; (2) perform a PCA or Multiple Factor Analysis with the environmental variables, creating linear combinations of highly correlated variables and then using scores to compute distance matrices between samples to be used as predictors (note that in this case predictors will be uncorrelated, in such a way that a bivariate Mantel test comparing community matrix with each predictor matrix will be satisfactory); (3) choose many groups of environmental variables randomly and perform many analyses, as suggested in the method of Clarke & Ainsworth (1993); and (4) construct a distance matrix for each environmental variable, and then define a minimum model using standard regression approaches, such as stepwise, to explain community structure. Of course, the last 2 alternatives involve a very large computational effort, especially alternative 3.

There is also another advantage in using the Mantel test as described above, related to the presence of qualitative environmental data, such as, for example, soil type. In a strict sense, canonical analyses do not work well with this kind of variable (making alternative 2 difficult). However, a specific environmental 'similarity' matrix for this variable (in alternative 4) would be constructed as a 'model matrix', in which the value 1.0 indicates that the 2 samples compared have the same type, or state, for the variable, and 0.0 indicates distinct states (Manly 1985). Used in this manner, the Mantel test works as a non-parametric multivariate analysis of variance with a randomization design. This type of qualitative environmental variable would also be combined with other environmental variables, after some standardization and choice of an adequate similarity coefficient (alternatives 1, 3 and 4).

So, we believe that the Mantel test is the best alternative for comparing environmental variables with multivariate community structure, due to its flexibility and multiple design possibilities, objectivity, powerful statistical basis and capacity for handling community structure in the original p-dimensional space without the need of clustering and ordination techniques and a posteriori (and sometimes ad hoc) interpretations based on them.

Acknowledgements. Our reseach programs in numerical ecology and population biology have been continuously supported by grants from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Coordenação para Aperfeiçoamento de Pessoal de Ensino Superior (CAPES).

LITERATURE CITED

Clarke KR, Ainsworth M (1993) A method of linking multivariate community structure to environmental variables. Mar Ecol Prog Ser 92:205-219

Crowley PH (1992) Resampling methods for computation-intensive data analysis in ecology and evolution. Ann Rev Ecol Syst 23:405-447

Dietz EJ (1983) Permutation tests for association between two distance matrices. Syst Zool 32:21-26

Dolédec S, Chessel D (1994) Co-inertia analysis: an alternative method for studying species-environmental relationships. Freshwat Biol 31:277-294

Fromentin JM, Ibanez F, Legendre P (1993) A phytosociological method for interpreting plankton data. Mar Ecol Prog Ser 93:285-306

Gauch HG (1987) Multivariate analysis in community ecology. Cambridge University Press, Cambridge

Gower JC (1966) Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53:325-524

Gower JC (1971) Statistical methods of comparing different multivariate analyses of the same data. In: Hodson FR, Kendall DG, Tautu P (eds) Mathematics in the archaeological and historical sciences. Edinburgh University Press, Edinburgh, p 138-149

Gower JC, Legendre P (1986) Metric and Euclidean properties of dissimilarity coefficients. J Classif 3:5-48

Harris RJ (1975) A primer of multivariate statistics. Academic Press, New York

Jackson DA, Somers KM (1989) Are probability estimates from the permutational model of Mantel's test stable? Can J Zool 67:766-769

James FC, McCulloch CE (1990) Multivariate analysis in ecology and systematics: panacea or Pandora's box. Ann Rev Ecol Syst 21:29-66

Johnson RA, Wichern DW (1992) Applied multivariate statistical analysis. Prentice-Hall, London

Krebs C (1989) Ecological methodology. Harper & Row, New York

Kruskal JB (1964a) Nonmetric multidimensional scaling: a numerical method. Psychometrika 29:115-129

Kruskal JB (1964b) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29:1-27

Lamont BB, Grant KJ (1979) A comparison of twenty-one measures of sites dissimilarity. In: Orloci L, Rao CR, Stiteler WM (eds) Multivariate methods in ecological work. International Co-operative Publishing House, Fairland, UK, p 101-126

Legendre L, Fortin MJ (1989) Spatial pattern and ecological analysis. Vegetatio 80:107-138

Legendre L, Legendre P (1983) Numerical ecology. Developments in environmental modelling, Vol 3. Elsevier Scientific Publishing Company, Amsterdam

Manly BFJ (1985) The statistics of natural selection. Chapman & Hall, London

Manly BFJ (1986a) Multivariate statistical methods: a primer. Chapman & Hall, London

Manly BFJ (1986b) Randomization and regression methods for testing for association with geographical, environmental and biological distance between populations. Res Popul Ecol 28:201-218

Manly BFJ (1991) Randomization and Monte Carlo methods in biology. Chapman & Hall, London

Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Res 27:209-220

Mantel N, Valand RS (1970) A technique of nonparametric multivariate analysis. Biometrics 26:547-558

Mielke PW (1978) Clarification and appropriate inferences for Mantel and Valand's nonparametric multivariate technique. Biometrics 34:277-282

Oden NL, Sokal RR (1992) An investigation of three-matrix permutation tests. J Classif 9:275-290

Palmer MW (1993) Putting things in even better order: the advantages of Canonical Correspondence Analysis. Ecology 74:2215-2320

Peet RK, Knox RG, Case JS, Allen RB (1988) Putting things in order: the advantages of detrended correspondence analysis. Am Nat 131:924-934

Rohlf FJ (1974) Methods for comparing classifications. Ann Rev Ecol Syst 5:101-113

Rohlf FJ (1982) Consensus indices for comparing classifications. Math Biosci 51:131-144

Rohlf FJ (1989) NTSYS: numerical taxonomy and multivariate analysis system. Exeter Softwares, New York

Smouse PE, Long JC, Sokal RR (1986) Multiple regression and correlation extensions of the Mantel test of matrix correspondence. Syst Zool 35:627-632

Sneath PHA, Sokal RR (1973) Numerical taxonomy. WH Freeman, San Francisco

Sokal RR (1979) Testing statistical significance of geographic variation patterns. Syst Zool 28:227-232

Sokal RR, Unnasch RS (1988) Geographic covariation of host and parasites: evidence from *Populus* and *Pemphigus*. Z Zool Syst Evolutionsforsch 26:73-88

Sokal RR, Wartenberg D (1981) Space and population structure. In: Griffith D, Mckinnon R (eds) Dynamic spatial models. Sijthoff and Noordhoff, Netherlands, p 186-213

Taylor CM, Gotelli NJ (1994) The macroecology of *Cyprinella*: correlates of phylogeny, body size and geographical range. Am Nat 144:549-569

ter Braak CJF (1986) Canonical Correspondence Analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67:1167-1179

ter Braak CJF (1987) The analysis of vegetation-environmental relationships by canonical correspondence analysis. Vegetatio 69:69-77

ter Braak CJF, Prentice IC (1988) The theory of gradient analysis. Adv Ecol Res 18:271-313

Wartenberg D, Ferson S, Rohlf FJ (1987) Putting things in order: a critique of detrended correspondence analysis. Am Nat 129:434-448

Washington HG (1984) Diversity, biotic and similarity indices. A review with special relevance to aquatic ecosystems. Water Res 18:653-694

Wolda H (1981) Similarity indices, samples size and diversity. Oecologia 50:296-302

This note was submitted to the editor

Manuscript first received: February 13, 1996
Revised version accepted: September 3, 1996