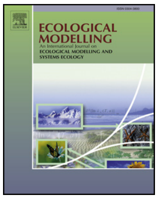
Ecological Modelling 383 (2018) 41–51

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel



Exploring multiple presence-absence data structures in ecology

János Podani[a,*], Péter Ódor[b,g], Simone Fattorini[c,d], Giovanni Strona[e], Jani Heino[f], Dénes Schmera[g,h]

[a] MTA-ELTE-MTM Ecology Research Group, Eötvös University, Pázmány P. s. 1/C, H-1117, Budapest, Hungary
[b] MTA Centre for Ecological Research, Institute of Ecology and Botany, Alkotmány u. 2-4, H-2163, Vácrátót, Hungary
[c] Department of Life, Health and Environmental Sciences, University of L'Aquila, L'Aquila, Italy
[d] CE3C – Centre for Ecology, Evolution and Environmental Changes/Azorean Biodiversity Group and University of Azores, Angra do Heroísmo, Portugal
[e] European Commission, Joint Research Centre, Directorate D - Sustainable Resources – Bio-Economy Unit, Via Enrico Fermi 274 9, 21027, Ispra, VA, Italy
[f] Finnish Environment Institute, Natural Environment Centre, Biodiversity, Paavo Havaksen Tie 3, FIN-90570, Oulu, Finland
[g] MTA Centre for Ecological Research, GINOP Sustainable Ecosystem Group, Klebelsberg K. u. 3, H-8237, Tihany, Hungary
[h] MTA Centre for Ecological Research, Balaton Limnological Institute, Klebelsberg K. u. 3, H-3237, Tihany, Hungary

ARTICLE INFO

Keywords:
Beta diversity
Classification
Comparison
Ordination
Similarity
Simplex diagram

ABSTRACT

Ecological studies may produce presence-absence data sets for different taxonomic groups, with varying spatial resolution and temporal coverage. Comparisons of these data are needed to extract meaningful information on the background ecological factors explaining community patterns, to improve our understanding of how beta diversity and its components vary among communities and biogeographical regions, and to reveal their possible implications for biodiversity conservation. A methodological difficulty is that the number of sampling units may be unequal: no method has been designed as yet to compare data matrices in such cases. The problem is solved by converting presence-absence data matrices to simplex plots based on the decomposition of Jaccard dissimilarity into species replacement and richness difference fractions used together with the complementary similarity function. Pairs of simplex plots representing different data matrices are then compared by quantifying, for each of them, the relative frequency of points in small, pre-defined subregions of the simplex, and then calculating a divergence function between the two frequency distributions. Given more than two data matrices, classification and ordination techniques may be used to obtain a synthetic and informative picture of metacommunity structure. We demonstrate the potential of our data analytical model by applying it to different case studies spanning different spatial scales and taxonomic levels (Mediterranean Island faunas; Finnish stream macroinvertebrate assemblages; Hungarian forest assemblages), and to a study of temporal changes in small islands (insect fauna in Florida). We conclude that, by accounting for various structural aspects simultaneously, the method permits a thorough ecological interpretation of presence-absence data. Furthermore, the examples illustrate succinctly how similarity, beta diversity and two of its additive components, species replacement and richness difference influence presence-absence patterns under different conditions.

1. Introduction

Community data derived from field surveys are routinely summarized in form of presence-absence matrices, with species (or other taxa) as rows, and study objects (e.g., sites, plots, localities, etc.) as columns. A given study may produce several data matrices from the same region which differ from one another in taxonomic coverage, spatial resolution, the time of sampling, or other ecologically meaningful factors. Similarly, meta-analyses attempting to summarize community level information from various and independent sources may also depend on different data matrices. In these cases, one is faced with the fundamental problem of comparing the *inherent structure of data matrices* with heterogeneous origin and properties. With 'structure' we refer here to non-random arrangement of presences/absences (i.e. 0-s and 1-s) in the matrix, revealing a tendency for grouping or nestedness, focusing, in particular, on structural patterns independent from the actual ordering of rows and columns (see Podani & Schmera 2011). Such comparisons are essential to understand variations among communities and biogeographical regions, the ecological factors explaining these patterns, and their possible implications for biodiversity conservation.

One possibility to tackle these issues is to perform classification or

---

[*] Corresponding author at: Department of Plant Systematics, Ecology and Theoretical Biology, Institute of Biology, Eötvös University, Pázmány P. s. 1.c, H-1117, Budapest, Hungary.
E-mail address: podani@ludens.elte.hu (J. Podani).

https://doi.org/10.1016/j.ecolmodel.2018.05.012
Received 8 February 2018; Received in revised form 10 May 2018; Accepted 10 May 2018

ordination on each data matrix and then to compare the resulting scatter plots, dendrograms or partitions. However, standard procedures available for this purpose can only be applied to cases where the number of study objects is the same in all the data sets under evaluation (Podani 2000). Alternatively, one may compare the data matrices directly, without multivariate analysis, but this methodology – in addition to equality in the number of objects – requires identical number of species as well. (Hubert & Golledge 1982; Zani 1986). That is, no universally applicable method has been developed as yet to compare the structure of data matrices that are unequal in size.

As a possible solution to this problem, we suggest a new analytical model that makes it possible to investigate multiple, heterogeneous datasets in a single framework. Essentially, the approach is based on the decomposition of Jaccard dissimilarities between pairs of objects into two additive components, namely species replacement (R) and richness difference (D), which, together with the complementary Jaccard similarity (S), are used to represent data structure as a point cloud in a ternary plot called SDR-simplex (Podani & Schmera 2011; Carvalho et al. 2012). Point clouds representing different data structures can then be compared on the basis of the relative frequencies of points (object pairs) in pre-defined subsections of the ternary plot. Since calculation of a frequency distribution is involved, we shall refer to this strategy as the *indirect* comparison of simplexes.

The approach is equally useful to situations where the matrices to be compared represent the same set of objects (for example, when a given set of objects is surveyed for different organism groups, or when a taxonomic group is examined in the same sites several times, to monitor temporal changes of community composition), even though in those cases comparisons of ordinations and classifications could also work. This is because decomposition of dissimilarity into additive terms allows separating the effects of major ecological driving forces – a possibility not available otherwise. Now, the simplex plots need not be partitioned; the shapes of point clouds can be *directly* compared by measuring the shift of the corresponding points in the two configurations.

Both indirect and direct comparisons may be performed on all possible pairs of matrices in a multiple dataset, yielding a dissimilarity matrix of SDR simplexes that can be then used in further analyses, such as classifications and ordinations. We emphasize here that this meta-analysis approach is more suited to exploratory analysis rather than to testing hypotheses. In this paper, we describe in detail the technical aspects of our method, and illustrate its potential in ecological research, by reporting results for both artificial examples and empirical case studies.

2. Computational steps

2.1. The SDR-simplex

Jaccard's (1901) similarity coefficient is one of the oldest and most commonly used resemblance functions, computed for any two objects *j* and *k* as:

$$s_{jk} = a / (a + b + c), \tag{1}$$

where a is the number of species occurring in both *j* and *k*, while b and c correspond to the number of species exclusive, respectively, to *j* and *k*. Its complement, Jaccard dissimilarity, is computed as:

$$\delta_{jk} = 1 - s_{jk} = (b + c) / (a + b + c). \tag{2}$$

Dissimilarity can be partitioned into two additive fractions (Podani & Schmera 2011; Carvalho et al. 2012):

$$\delta_{jk} = d_{jk} + r_{jk} = | b–c | / (a + b + c) + 2 \min\{b,c\} / (a + b + c), \tag{3}$$

where $d_{jk} = | b–c | / (a + b + c)$ is the relative richness difference, while $r_{jk} = 2 \min\{b,c\} / (a + b + c)$ is the relative species replacement with respect to objects *j* and *k*. In the latter, the numerator is the maximum fraction of the so-called species turnover, which is equally shared by *j* and *k*. Since $s_{jk} + d_{jk} + r_{jk} = 1$, these three quantities may be used to define the relative position of the point representing object pair *jk* with respect to the three vertices (S–Similarity, D–richness Difference and R–species Replacement) of an equilateral triangle, the so-called SDR-simplex diagram (Podani & Schmera 2011). In the SDR-simplex, the distance of each point from a given vertex is inversely proportional to the corresponding fraction, that is, S, D or R. Similar ternary plots have been used in ecology as illustrations of C-S-R strategies of plants (Grime 1977), of feeding habits of fish (Fig. 6.9 in Stoffels 2013), and are even more widely used in population genetics (commonly referred to as "de Finetti diagram") to represent the genotype frequencies of diploid populations for a biallelic locus (Edwards 2000), and in geology to classify rocks and minerals on the basis of their fractional composition (Streckeisen 1976).

Let us first demonstrate the procedure for a pair of hypothetical objects *j* and *k* containing a total of 12 species with different nonzero values of a, b and c. If the objects have many species in common (a = 8), and species replacement and richness differences are equal (2 min{b,c} = | b–c | = 2), then the point representing this pair of objects in the ternary plot is positioned close to the S vertex, and with equal distance from D and R (Fig. 1a). If richness difference is high (| b–c | = 9) and similarity and replacement are the same (a = 2 min {b,c} = 2), then the point moves close to the D vertex (Fig. 1b). Analogously, if species replacement is the dominating phenomenon, with 4 species being replaced by other 4 (2min{b,c} = 8), and the two objects sharing only 2 species, with a richness difference of 2 (a = | b–c | = 2), the point in the ternary plot is positioned close to the R vertex, and with equal distance from D and S (Fig. 1c). When the three components are equal (a = 2 min{b,c} = | b–c| = 4), the corresponding point will fall onto the center of the triangle (Fig. 1d).

In a data matrix **X** containing m objects, the possible number of pairwise comparisons would be $w = (m^2 – m) / 2$, each corresponding to a point in the simplex. Notably, the shape of the point cloud in a simplex is unaffected by the actual arrangement of rows and columns in the matrix. "Extreme" structural patterns produce clear distributions of points in the triangle. If compositional similarity is high for all pairs, the point cloud will be near the S corner. When the objects have extreme richness differences, with low replacement and similarity, the points will be close to the D vertex. In cases when richness is similar but similarity is low, the points will be in the upper third of the diagram. These cases are extensions of the two-object situations explained above and are not illustrated. However, there are further noted examples in which two of the three components contribute approximately equally to data structure, whereas the third is zero. Maximum beta diversity (antinestedness) in the data (with $s_{jk} = 0$ for all $j \neq k$), makes all points fall onto the left (D-R) side of the triangle (Fig. 1e), while maximum nestedness of objects (with $r_{jk} = 0$ for all pairs) forces all points to the bottom (D-S) side (Fig. 1f). In case of a perfect gradient (when the species richness is constant, the same number of species are lost and gained at each sampled step along that gradient, and $d_{jk} = 0$ for all pairs) all points are distributed on the S-R side (Fig. 1g). See Podani & Schmera (2011), for further examples of structural patterns and their simplex representations. The position of the centroid of the point cloud (calculated as the means of the s_{jk}, d_{jk} and r_{jk} values) will be used in a synthetic measure to compare the structure of comparable plots. Furthermore, these means multiplied by 100 quantify the percentage contributions of the three fractions to community pattern. In addition to these contributions, it is also useful to consider the percentage of presence scores in the data matrix, i.e., matrix fill, denoted here by *q*.

2.2. Indirect comparison of two SDR-simplexes for different sets of objects

Assume that we have another data matrix, **Y**, with p objects.