CONCEPTS & SYNTHESIS
EMPHASIZING NEW IDEAS TO STIMULATE RESEARCH IN ECOLOGY

Ecological Monographs, 75(4), 2005, pp. 435–450

ANALYZING BETA DIVERSITY: PARTITIONING THE SPATIAL VARIATION OF COMMUNITY COMPOSITION DATA

PIERRE LEGENDRE,[1] DANIEL BORCARD, AND PEDRO R. PERES-NETO[2]

Département de sciences biologiques, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Québec H3C 3J7 Canada

Abstract. Robert H. Whittaker defined beta diversity as the variation in species composition among sites in a geographic area. Beta diversity is a key concept for understanding the functioning of ecosystems, for the conservation of biodiversity, and for ecosystem management. This paper explains how hypotheses about the origin of beta diversity can be tested by partitioning the spatial variation of community composition data (presence–absence or abundance data) with respect to environmental variables and spatial base functions. We compare two statistical methods to accomplish that. The sum-of-squares of a community composition data table, which is one possible measure of beta diversity, is correctly partitioned by canonical ordination; hence, canonical partitioning produces correct estimates of the different portions of community composition variation. In recent years, several authors interested in the variation in community composition among sites (beta diversity) have used another method, variation partitioning on distance matrices (Mantel approach). Their results led us to compare the two partitioning approaches, using simulated data generated under hypotheses about the variation of community composition among sites. The theoretical developments and simulation results led to the following observations: (1) the variance of a community composition table is a measure of beta diversity. (2) The variance of a dissimilarity matrix among sites is not the variance of the community composition table nor a measure of beta diversity; hence, partitioning on distance matrices should not be used to study the variation in community composition among sites. (3) In all of our simulations, partitioning on distance matrices underestimated the amount of variation in community composition explained by the raw-data approach, and (4) the tests of significance had less power than the tests of canonical ordination. Hence, the proper statistical procedure for partitioning the spatial variation of community composition data among environmental and spatial components, and for testing hypotheses about the origin and maintenance of variation in community composition among sites, is canonical partitioning. The Mantel approach is appropriate for testing other hypotheses, such as the variation in beta diversity among groups of sites. Regression on distance matrices is also appropriate for fitting models to similarity decay plots.

Key words: beta diversity; canonical ordination; community composition; Mantel test; PCNM analysis; regression on distance matrices; simulation study; spatial variation; variation partitioning.

INTRODUCTION

The variation in species composition among sites, within a region of interest, was termed ''beta diversity'' by Whittaker (1960, 1972). Proper management of ecosystems, which are often simply seen by tenants of neoliberalism as resources for the industry, requires that we understand the processes by which beta diversity is created and maintained. If beta diversity is entirely the result of contemporary and historical random processes, we can take resources anywhere without adverse effects as long as we are not depleting them. If it is not, we have to preserve the spatial organization or the species–environment relationships necessary for nature to recreate and maintain beta diversity.

The main current hypotheses about the origin of beta diversity are as follows:

Manuscript received 6 April 2005; accepted 25 April 2005; final version received 23 May 2005. Corresponding Editor: N. G. Yoccoz.

[1] E-mail: Pierre.Legendre@umontreal.ca

[2] Present address: Department of Biology, University of Regina, Regina, Saskatchewan S4S 0A2 Canada.

1) *Species composition is uniform over large areas.* This hypothesis, which plays the role of a null model, emphasizes the role of *biological interactions.* It suggests that communities are dominated by a limited suite of competitively superior species (Pitman et al. 1999, 2001); beta diversity is small.

2) *Species composition fluctuates in a random, autocorrelated way.* This hypothesis emphasizes spatially limited *dispersal history.* Models derived from neutral theory state that all species are demographically and competitively equal. Differences are created through spatially limited dispersal of species drawn at random from a metacommunity, plus possibly the appearance of newly evolved species in different areas. Neutral models differ in the details of the mechanisms that they invoke (Bell 2001, Hubbell 2001, He 2005).

3) *Species distributions are related to environmental conditions.* This hypothesis emphasizes *environmental control.* Landscapes are mosaics where species composition is controlled by environmental site characteristics (Whittaker 1956, Bray and Curtis 1957, Hutchinson 1957, Gentry 1988, Tuomisto et al. 1995).

Testing these hypotheses is important for understanding the functioning of ecosystems, for the conservation of biodiversity, and for ecosystem management. Regarding the establishment of natural reserves, e.g., hypothesis 1 implies that all parts of the ecosystem are equivalent. Reserves can be located anywhere. Hypothesis 2 implies that different parts of the ecosystem may, for historical reasons, sustain different species compositions, although these parts are environmentally equivalent. Portions of space supporting the different species compositions should be preserved. Reserves must be large, allowing the dynamics to go on without many species going extinct. Hypothesis 3 implies that all parts of the ecosystem are not equivalent. Reserves must represent the different types of habitat and each portion must be of sufficient size to be sustainable. The parts representing favorable dispersion routes must be especially preserved.

When studying the origin of beta diversity, one must consider different hypotheses and answer the following questions. (1) Is the variation in species composition among sites *random,* i.e., devoid of significant spatial pattern? A positive answer will support hypothesis 1. (2) Is there *significant spatial patchiness* (different from random) in the distributions of species? A positive answer will support hypothesis 2, and possibly hypothesis 3 if the environmental variables influencing species distributions are spatially structured. (3) Can the *environmental variables* explain a significant proportion of the community composition variation? A positive answer will support hypothesis 3, which is compatible with hypothesis 2.

Empirical data must be used to determine the likelihood of each hypothesis in different systems and at different spatial scales. To test these hypotheses, the variation of community composition at many sites must be analyzed to determine if *significant spatial patterns* (different from random) are present in the data and if the *environmental variables* explain a significant proportion of the community composition variation. In this contribution, we explain how hypotheses about the origin of beta diversity can be tested by partitioning the spatial variation of community composition (presence–absence or abundance data) with respect to environmental variables and spatial base functions. Next we set out to assess the appropriateness and robustness of two partitioning approaches used by researchers to assess the likelihood of different mechanisms structuring beta diversity.

Beta Diversity

Alpha (α) diversity is the diversity in species at individual sites (e.g., plots, quadrats), or the variance in the species identity of the individuals observed at a site. A monoculture, for example, has the lowest possible alpha diversity because there is no variance in the species identity of individuals. It is measured either by the number of species present at the site (species richness), or by some other function (index) that takes into account the relative frequencies of the species. The most widely used of such indices are the Shannon (1948) (or Shannon-Weaver, or Shannon-Wiener) index of diversity and measures based on Simpson's (1949) concentration λ, which are often referred to as Simpson's indices: either $(1 - \lambda)$ (Greenberg 1956) or $1/\lambda$ (Hill 1973). The form $(1 - \lambda)$ is to be preferred because, when applied to combined sites (γ diversity, next paragraph), that measure cannot be smaller than the mean of the α indices for the sites that have been combined (Lande 1996).

Gamma (γ) diversity is that of the whole region of interest for the study. It is usually measured by pooling the observations from a sample (in the statistical sense), i.e., a large number of sites from the area, except in those rare instances in which the community composition of the entire area is known. Gamma diversity is measured using the same indices as alpha diversity.

Beta (β) diversity is the variation in species composition among sites in the geographic area of interest. If the variation in community composition is random, and accompanied by biotic processes (e.g., reproduction) that generate spatial autocorrelation, a gradient in species composition may appear and beta diversity can be interpreted in terms of the rate of change, or turnover, in species composition along that gradient. If differentiation among sites is due to environmental factors, beta diversity should be analyzed with respect to the hypothesized forcing variables. In many ecosystems, beta diversity may be caused concurrently by these two processes in different proportions. In two seminal papers, Whittaker (1960, 1972) showed that beta diversity could be measured from either presence–absence or quantitative species abundance data. Ecologists actually use both to study beta diversity. Some

only refer to presence–absence when they talk about the rate of species replacement, or *turnover*, along ecological gradients. In the ordination literature, however, ecologists most often use species abundances to study turnover rates by reference to the appearance and disappearance of species with unimodal distributions along a gradient.

If one is interested in a global interpretation of beta species diversity within a study area, one can turn to the methods described in the section *Analyzing beta diversity*. When the goal is to compare different areas, or different taxonomic groups within the same area, one may be interested in using a single number summarizing beta diversity. The total sum of squares in the species composition data at the sampling sites, SS(**Y**), is one such number. The total variance in the data table, $\text{Var}(\mathbf{Y}) = \text{SS}(\mathbf{Y})/(n - 1)$, would be better to compare areas with different numbers of study sites n. The beta diversity indices proposed by Whittaker and others are other such measures.

The first method for obtaining a global measure of beta diversity from species presence–absence data, proposed by Whittaker (1960, 1972), is to compute the ratio of two diversity indices: $\beta = S/\bar{\alpha}$, where S is the number of species in a composite community composition vector representing the area of interest, whereas $\bar{\alpha}$ is the mean number of species observed at the original sites. This is a multiplicative approach. S represents gamma diversity in this equation. This measure of beta diversity describes how many more species are present in the whole region than at an average site, and uses that value as the measure of beta diversity.

An alternative additive approach, which has been present in the literature since MacArthur et al. (1966), Levins (1968), and Allan (1975), has been revived by Lande (1996) and widely adopted since then (Veech et al. 2002). In that approach,

$$D_{\text{T}} = D_{\text{among}} + \bar{D}_{\text{within}}$$

where D_{T} is the total (gamma) diversity. This approach can be applied to species richness as well as Shannon information and Simpson diversity $D = (1 - \lambda)$; see Lande (1996) for details. Because diversities are variances, one easily recognizes an analysis of variance approach in that equation, but the analogy ends there. The partitioning method described in *Partitioning community composition variation among groups of explanatory variables* and used in our numerical simulations is another form of additive partitioning (see Appendix A). Interestingly, Whittaker (1960, 1972) had suggested that beta diversity could also be quantified by dissimilarity matrices among sites, using either presence–absence data (computing the one-complement of the Jaccard or Sørensen coefficients of similarity or the index $\beta = S/\bar{\alpha}$ for pairs of sites), or species composition data using the one-complement of Whittaker's (1952) index of association or the Steinhaus coefficient of similarity (the latter has been redescribed, in dissimilarity form, by Odum [1950], and by Bray and Curtis [1957]). In these cases, Whittaker (1972) suggested that the *mean* (not the variance) of the dissimilarities among sites should be used as the measure of beta differentiation: ''the mean CC [Jaccard's coefficient of community] for samples of a set compared with one another [. . .] is one expression [of] their relative dissimilarity, or beta differentiation'' (Whittaker 1972:233). This idea had been suggested by Koch (1957: Table III and analysis in the text), who was looking for an *index of biotal dispersity*. Whittaker thus recognized that the dissimilarities are themselves measures of the differentiation between or among sites. In Eq. 1, we will study in more detail the relationship between dissimilarity matrices and the total sum of squares of community composition data matrices, used in variation partitioning to assess the likelihood of hypotheses 2 and 3.

Another approach to measuring α and β diversity is through the methods of ordination. Interrelationships between ordination techniques and diversity indices have been known since Gauch and Whittaker (1972). They were recently reviewed and synthesized by Pélissier et al. (2003).

(1) We will first review, in the next section, the methodological developments that led to the present dilemma between a ''raw-data'' and a ''distance'' approach. (2) We will then show that the total among-site variance of raw species data tables can be estimated by analysis of dissimilarity matrices. Thus, measures of beta diversity can be obtained from dissimilarity matrices computed in various ways and corresponding to various ecological assumptions. (3) We will briefly discuss how beta diversity can be analyzed by clustering or ordination. (4) In the sequel, we will discuss how hypotheses about the origin of beta diversity can be tested by canonical analysis of raw species presence–absence and abundance data tables. (5) We will show how the variance of a species composition data table can be partitioned among groups of explanatory variables. (6) We will then show that while the mean of the dissimilarities provides a measure of beta diversity, the *variance* of the dissimilarities does not. Hence, a decomposition of the variance of the dissimilarity matrix among different sets of explanatory variables cannot help us to test hypotheses, stated in terms of the raw community composition data, about the origin of beta diversity. (7) Finally, we will briefly discuss the analysis of (dis)similarity matrices based on community composition data.

Analyzing Beta Diversity

Analyze raw data or distance matrices?

The choice between the two analytical approaches requires than the researcher be aware of the level of abstraction addressed. (1) *Variation in species identity within communities*: at the basis of all communities are

species and their abundances. Studying variation in species identity of individuals at a given site is studying alpha diversity. (2) *Variation in community composition among sites*: the number, identity, and abundance of species define a community composition, which may vary among sites. Studying variation of community composition among sites is studying beta diversity and may be done using the raw-data approach. (3) *Variation in beta diversity among groups of sites*: if measured among pairs of sites, beta diversity may vary from pair to pair. Therefore, studying how and why beta diversity varies among regions or groups of sites pertains to a third level of abstraction, which must not be confused with the previous one. The hypotheses arising at this level may be addressed using either the distance (i.e., Mantel) approach or the tests of significance recently proposed by Kiflawi and Spencer (2004). This paper is devoted to the methods that are appropriate for testing hypotheses that concern the variation in community composition among sites, i.e., those that pertain to the second level of abstraction.

Because of the ease with which spatial relationships among sampling sites can be incorporated into the analysis in the form of a geographic distance matrix, Legendre and Troussellier (1988) proposed using partial Mantel analysis to test hypotheses about the relative influence of the spatial structure and environmental variables on ecological response variables. Population geneticists had already successfully experimented with the Mantel approach for similar problems (e.g., Smouse et al. 1986). A few years later, Borcard et al. (1992) proposed using variation partitioning through partial canonical analysis (redundancy analysis [RDA] or canonical correspondence analysis [CCA]) to partition community variation among environmental and spatial components. At that time, canonical analysis was limited to the Euclidean or chi-square distances for response data. The Mantel approach offered more flexibility, allowing the use of other types of distance functions such as Jaccard or Steinhaus/Bray-Curtis. Also, the canonical approach was limited to analyzing spatial gradients or broadscale spatial structures modeled by a polynomial trend-surface function of the geographic coordinates. The Mantel approach was limited to statistical testing only, whereas canonical analysis provided estimates of the contributions of the response and explanatory variables to the canonical relationship, as well as the visual analysis of patterns in ordination biplots. In 1993 and 1998, P. Legendre concluded that both methods had been useful in enriching our understanding of spatial processes occurring in ecosystems (Legendre 1993, Legendre and Legendre 1998: Section 13.6). In the absence of statistical reasons for choosing one or the other, he recommended using both approaches until comparative studies provided criteria for choosing one or the other. Fortin and Gurevitch (1993) also recommended the Mantel approach for vegetation studies.

Mantel correlations between distance matrices are known to be much smaller in absolute value than regular correlation coefficients computed on the same raw data (Dutilleul et al. 2000, Legendre 2000). This is thus also the case when comparing coefficients of determination in regression on distance matrices and in multiple regression or canonical analysis. Canonical analysis partitions the variation in the species abundance data and therefore explains a greater amount of the total variation than does the Mantel test, which partitions the variation in pairwise dissimilarities among sites. Legendre (2000: Table II) reported simulations showing that, for two simple variables, the Pearson correlation, which is the simplest form of the raw-data approach, had more power than the Mantel test computed on the same data. That paper showed that the Mantel test was inappropriate for testing hypotheses concerning variation of the raw data, although it may be appropriate for hypotheses concerning variation of distances. Using numerical simulations, we will confirm the last point for multivariate data and compare the statistical power of the two procedures for partitioning the variation of raw data. We will also present our most recent thoughts about differentiating the raw-data and Mantel approaches.

Two handicaps of the raw-data approach have been lifted in recent years: Legendre and Gallagher (2001) showed how to transform community composition data in such a way that distances that are of interest for community ecology (e.g., the chord, chi-square, and Hellinger distances) would be preserved during canonical redundancy analysis. Borcard and Legendre (2002), Borcard et al. (2004), and S. Dray, P. Legendre, and P. Peres-Neto (*unpublished manuscript*) developed more complex forms of spatial representation (PCNM analysis [principal coordinates of neighbor matrices] and related distance-based eigenvector maps) that allow ecologists to model spatial structures at multiple spatial scales and incorporate these representations in canonical analyses to answer the questions mentioned in the *Introduction*.

Many authors have used either partial Mantel tests (Smouse et al. 1986) or multiple regression on distance matrices (Legendre et al. [1994]; computer program, Casgrain [2001]), with the objective of analyzing the spatial variation of community composition among sites, while in fact these methods should be limited to the study of the variation in beta diversity among groups of sites. Here are examples from the recent literature in which authors used a Mantel approach, often citing Legendre and Legendre (1998) or Fortin and Gurevitch (1993) as references, although they declared that the purpose of their study was the analysis of the variation in community composition among sites. Examples of vegetation analyses include: Svenning (1999; tropical palms), Potts et al. (2002; tropical trees), Phillips et al. (2003; tropical trees), Reynolds and Houle (2003; *Aster*). Examples of animal analyses include:

Parris and McCarthy (1999; frog assemblages), Somerfield and Gage (2000; macrobenthos communities), Pusey et al. (2000; stream fish), Decaens and Rossi (2001; earthworm communities), Orwig et al. (2002; homopteran insects), Spencer et al. (2002; invertebrate communities), Williams et al. (2003; fish and macroinvertebrates), Mulder et al. (2003; soil nematode communities). The present paper shows that canonical analyses of the species abundance data would have provided more powerful tests of significance than analyses based on distance matrices. They also would have produced estimates of the contributions of the response and explanatory variables to the canonical relationship, as well as bi- or triplots to illustrate the relationships; the distance approach does not provide estimates of the contribution of individual variables to the overall Mantel correlation.

Authors have also used Mantel tests to compare seed banks to actual vegetation; examples are Erkkilä (1998) and Jutila (2002). In the same vein, there are papers that used Mantel tests to study the concordance (or congruence) of communities, comparing the spatial patterns of two communities observed at a series of sites. Examples are: Paszkowski and Tonn (2000; fish and aquatic birds) and Su et al. (2004; birds, butterflies and vascular plants). Researchers would have obtained more powerful tests using either co-correspondence analysis (ter Braak and Schaffers 2004), or canonical correlation analysis of transformed community composition data.

Total beta diversity measured from a dissimilarity matrix

We saw, in the section *Beta diversity*, that beta diversity is the variation in species composition among sites in a geographic area. The variance of the species composition data table thus provides one possible measure of the total beta diversity in that area, which is the total sum of squares (SS) of that table.

From Whittaker (1972), we learned that a dissimilarity matrix (**D**) among sites within a study area, based on community composition data, is also, in itself, an expression of the beta diversity of that area. We will see in the next paragraphs that these two notions of beta diversity are mathematically related and that SS can be obtained from a dissimilarity matrix **D**. It is possible to compute **D** using any dissimilarity function deemed appropriate to express the community composition relationships among sites. Because this is a key point in our argument, we will carefully walk the readers through the steps of the reasoning.

The total sum of squares (SS) of a multivariate rectangular data table $\mathbf{Y}(n \times p)$, such as a species composition table involving n sites and p species, is the sum, over all species, of the squared deviations from the mean of each species. Dividing SS by $(n - 1)$, where n is the number of sites, would produce the classical unbiased estimate of the total variance of **Y**. For simplicity, the development that follows is written in terms of sums of squares instead of variances.

SS can also be computed as the sum of squared Euclidean distances (D^2_{hi}) among the row vectors of table **Y** (sites) divided by the number of sites (n):

$$\mathrm{SS}(\mathbf{Y}) = \sum_{i=1}^{n} \sum_{j=1}^{p} (y_{ij} - \bar{y}_j)^2 = \frac{1}{n} \sum_{h=1}^{n-1} \sum_{i=h+1}^{n} D^2_{hi}. \quad (1)$$

This well-known relationship is presented as Theorem 1 in Legendre and Anderson (1999: Appendix B); it is also illustrated in Legendre and Legendre (1998: Fig. 8.18).

Consider any real-data example involving two sites only, each one represented by a vector of species presence or abundance data. First, compute the total sum of squares in that $2 \times p$ data table after centering the species on their means (for the two sites only). Compute a 2×2 dissimilarity matrix among these sites using the Euclidean distance formula. The total sum of squares in the $2 \times p$ data table is equal to the squared dissimilarity divided by 2 (2 is the number of sites). This shows that a single dissimilarity measure contains the information relative to the sum of squares between two sites: the sum of squares of the original species data can be obtained by calculating $D^2/2$. *Lessons learned:* (1) the variance between two sites can be obtained from the dissimilarity D computed between them. (2) A dissimilarity matrix among sites within an area, based on community composition data **Y** and computed using the Euclidean distance formula, is an expression of the total sum of squares of table **Y**, and thus the beta diversity of that area.

Community ecologists often transform community composition data in a way deemed appropriate for their study, such as square-root, double square-root, log, or one of the transformations described by Legendre and Gallagher (2001). These transformations may be useful in a variety of cases such as weighting down the importance of very abundant species, making the data more linear in relation to exogenous predictors, or making the analysis invariant with respect to double absences. A dissimilarity matrix among sites obtained by computing the Euclidean distance on such transformed data is thus an expression of the beta diversity in the area after modeling the community composition data in some appropriate way. The dissimilarity is no longer the Euclidean distance for the original data, but the Euclidean distance computed on the transformed data.

Likewise, community ecologists often use dissimilarity functions developed to model community composition relationships in other ways than the Euclidean distance does. Examples of suitable functions are the chord, Hellinger, chi-square, and Bray-Curtis dissimilarities, and the one-complement of Whittaker's index of association. These dissimilarity functions can be seen as transformations of the original community composition data, because a principal coordinate anal-

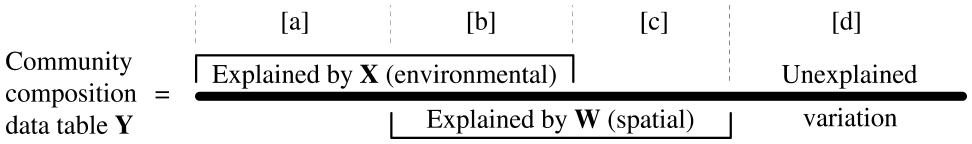

FIG. 1. Partition of the variation of a response matrix **Y** between environmental (matrix **X**) and spatial (matrix **W**) explanatory variables. The bold horizontal line, which corresponds to 100% of the variation in **Y**, is divided into four fractions labeled [a] to [d]. The figure is adapted from Borcard et al. (1992) and Legendre (1993).

ysis of the dissimilarity matrix produces a rectangular data table that can replace the original data in further analyses (Legendre and Anderson 1999). Some of these dissimilarities (chord, chi-square, Hellinger) can be obtained by directly transforming the original data table and then computing the Euclidean distance on the transformed data (Legendre and Gallagher 2001). Using the transformed data directly in canonical analyses permits one to relate the species data table to the explanatory variables of beta diversity. *Lessons learned:* (1) a dissimilarity matrix among sites, computed from one of the functions commonly used by community ecologists, is an expression of the beta diversity in the area under study. (2) For a dissimilarity matrix computed using a function that is considered appropriate for community composition data, the value $1/n\ \Sigma D_{ij}^2$ is equal to SS(**Y**), the sum of squares of the transformed community composition data table, which is a measure of beta diversity.

Analyzing beta diversity by clustering or ordination

Besides the estimation of the total sum of squares, the variation in community composition among sites (beta diversity) can be analyzed in various ways using the dissimilarity matrix **D**. For example, one may describe it by looking for geographically compact clusters of similar sites separated by discontinuities, using clustering or partitioning methods. One may also look for gradients using ordination methods.

Testing hypotheses about the origin of beta diversity using community composition data

Hypotheses about the origin of beta diversity, such as those outlined in the *Introduction*, can be tested by using standard forms of linear modeling such as multivariate ANOVA or canonical analysis. Because the hypotheses concern the origin of the variation in community composition among sites, tests of significance must be carried out on the original (or transformed) community composition data table, not on distance matrices. In some instances, one may use a proxy table obtained by principal coordinate analysis of an appropriate form of dissimilarity matrix (db-RDA approach; Legendre and Anderson 1999). The main forms of analysis used by ecologists are canonical redundancy analysis (Rao [1964], who called this method ''principal components of instrumental variables'') and canonical correspondence analysis (ter Braak 1986, 1987*a*, *b*).

As stated in the *Introduction*, hypotheses about the processes that create or maintain beta diversity invoke either community dynamics creating spatial patterns, or environmental control, or both. These questions are particularly well suited for canonical analysis. Community composition is the response data table (**Y**) in the analysis, whereas a table of environmental variables and/or some representation of the spatial relationships forms the explanatory table (**X**). The representation of spatial relationships in such analyses is described in the next subsection. As in subsection 1, community composition can be analyzed as presence–absence or abundance data, raw or transformed.

Partitioning community composition variation among groups of explanatory variables

Based on the general statistical method of partitioning of the variation of a response variable among two or more explanatory variables, called ''commonality analysis'' by Kerlinger and Pedhazur (1973: Chapter 11) for the univariate case, the method of partitioning the variation of multivariate community composition data among environmental and spatial components, through canonical analysis, was developed and described by Borcard et al. (1992, 2004), Borcard and Legendre (1994, 2002), Legendre and Legendre (1998), and Méot et al. (1998). That form of statistical analysis allows users to decompose the variation found in a multivariate table of response variables (e.g., community composition data) as a function of a set of environmental variables (in a broad sense, i.e., variables corresponding to hypotheses related to the environmental control model, the biotic control model, or describing historical dynamics) and a set of spatial variables constructed from the geographic coordinates of the sites. One can calculate [a + b] how much of the variation is explained by the environmental variables, [b + c] how much is spatially structured, [b] how much of the environmentally explained variation is spatially structured, and [d] how much remains unexplained (Fig. 1).

There are different ways of representing geographic information in the calculations, e.g., raw geographic coordinates of the sampling sites if one wants to fit a flat (planar) trend surface response to the data, or a polynomial function of the geographic coordinates in order to model a wavy or saddle-shaped trend surface (Legendre 1990). Other types of spatial structures can be modeled, e.g., a river network connecting lakes (Magnan et al. 1994). A recent development for modeling the spatial variation at multiple scales is PCNM

analysis, described by Borcard and Legendre (2002) and Borcard et al. (2004). Other forms of distance-based eigenvector maps (DBEM) related to PCNM, based on different types of spatial weighting matrices, are described in S. Dray, P. Legendre, and Peres-Neto (*unpublished manuscript*). Another interesting development is the suggestion by Wagner (2004) to take up the fractions of variation obtained by Borcard et al. (1992) partitioning, and reanalyze them by distance classes through a multivariate variogram to identify patterns corresponding to particular scales. That method allows one to further refine and test the hypotheses stated in our *Introduction*. Is the among-sites beta diversity spatially structured? Is the among-sites beta diversity related to environmental variables in a scale-dependent manner? Are the residuals spatially autocorrelated and, if so, can this be interpreted as the signature of biotic processes or as the indication that important environmental variables are missing?

The method of multivariate variation partitioning can be directly extended to several explanatory data tables, as shown by Quinghong and Bråkenhielm (1995), Anderson and Gribble (1998), Pozzi and Borcard (2001), and Økland (2003). If the explanatory table of spatial coordinates is replaced by a table of dummy variables representing geographic regions, for example, the method becomes partial multivariate analysis of variance (Legendre 1993: Fig. 3) and allows ecologists to obtain estimates of within- and among-region diversity, in the spirit of the additive partitioning of Lande (1996).

Appendix A shows that canonical variation partitioning provides a correct partitioning of the variation of a response data table **Y**. Canonical partitioning also provides tests of significance of all the fractions of variation that can be estimated by a canonical model, estimates of the contributions of individual explanatory variables or groups of them, and plots displaying the relationship between species, sites, and explanatory variables.

Partitioning the variation of dissimilarity matrices

Some of the papers on beta diversity, listed as examples in the section *Analyze raw data or distance matrices,* have used variation partitioning based on regression on distance matrices, a method (and computer program) developed by Legendre et al. (1994) for phylogenetic analysis. (Analyses involving correlation or regression on distance matrices are often globally referred to as the Mantel approach because the scalar product of two **D** matrices and regression on such matrices was first suggested by Mantel [1967].) The variation among sites (beta diversity) is represented in the analysis by a community composition *distance matrix*, which is analyzed by multiple regression against *distance matrices* representing environmental variation and geographic relationships. In the multiple regression, the explanatory distance matrices explain a portion of the variation of the community composition distance matrix (R^2).

This method of analysis is inappropriate when the hypothesis to be tested concerns the raw data collected to study the origin of variation in species composition among sites (beta diversity), because it partitions the variation of a dissimilarity matrix SS(**D**), not the variation SS(**Y**) of the community composition data table. Although the *mean* of the squared dissimilarities provides a measure of beta diversity (Eq. 1 and upper portion of Eq. 2), the *variation* (sum of squares) of the dissimilarities (lower portion of Eq. 2) does not, because it is not equal to, nor is it a simple function of, the variation of the original data, SS(**Y**):

$$\mathrm{SS}(\mathbf{Y}) = \frac{1}{n}\sum_{h=1}^{n-1}\sum_{i=h+1}^{n} D_{hi}^2 \neq$$

$$\mathrm{SS}(\mathbf{D}) = \sum_{h=1}^{n-1}\sum_{i=h+1}^{n} (D_{hi} - \bar{D})^2$$

$$= \sum_{h=1}^{n-1}\sum_{i=h+1}^{n} D_{hi}^2 - \frac{\left(\sum\sum D_{hi}\right)^2}{n(n-1)/2}. \qquad (2)$$

Consider the numbers 1 to 10, for example. Their total sum of squares is 82.5 (Eq. 1 and upper portion of Eq. 2). Compute an Euclidean distance matrix **D**, SS(**Y**), among these 10 numbers: the sum of squares of the distances in the upper-triangular portion of matrix **D**, SS(**D**), is 220 (Eq. 2, lower portion). Hence, a decomposition of the variance of the dissimilarity matrix among two or several explanatory tables represented by dissimilarity matrices cannot help us to understand the causes of the variation of community composition (beta diversity) across an area.

The variation in matrix **D** can also be lower than the variation in **Y**. It can actually be null, as shown in Fig. 2; this is an extreme example showing that the variation in a **D** matrix bears no relationship to the variation of the original or Jaccard-transformed data. There is clearly beta diversity among the four sites in that example. We can measure it either by computing the SS of the raw data (Fig. 2c), or by applying the right-hand side of Eq. 1 to the Jaccard distance matrix (Fig. 2d). The variation among the distances in matrix **D** is zero, however (Fig. 2e). This is clearly an incorrect measure of the beta diversity of the species data. This is not presented as an example of a real ecological situation, but simply as an illustration of a major difference in calculation results between the two methods.

Analysis of community composition (dis)similarity matrices

We are not suggesting that the similarity/distance matrix approach is always inappropriate in the context of beta diversity studies. Indeed, the causes of the variation in a distance matrix offer opportunities to formulate and test interesting ecological hypotheses con-

a) Data

	Sp.1	Sp.2	Sp.3	Sp.4	Sp.5
Site 1	1	1	0	0	0
Site 2	1	0	1	0	0
Site 3	1	0	0	1	0
Site 4	1	0	0	0	1

b) **D** = [1 – Jaccard similarity]

	Site 1	Site 2	Site 3	Site 4
Site 1	0	0.667	0.667	0.667
Site 2	0.667	0	0.667	0.667
Site 3	0.667	0.667	0	0.667
Site 4	0.667	0.667	0.667	0

c) Total SS of raw data = 3.000

d) Total SS of Jaccard-transformed data

$$= \frac{1}{n} \sum_{h=1}^{n-1} \sum_{i=h+1}^{n} D_{hi}^2 = 0.667$$

e) Total SS of distances in the upper triangle = 0

FIG. 2. (a) Community composition data table and (b) derived distance matrix (one-complements of Jaccard similarities). (c) Total variation (sum of squares, SS) in table (a). (d) Total variation computed from matrix (b) for the Jaccard-transformed data (a). (e) Total SS of the distances in the upper triangle of the distance matrix in (b).

cerning the origin of the variation in beta diversity among groups of sites (level-3 questions; see *Analyzing beta diversity*). For example, Tuomisto et al. (2003*b*) looked for areas of high variability in floristic distances and tried to interpret them. Another example is Hubbell's neutral model of biodiversity, which predicts that the shape of the relationship between community similarity (based on community composition, i.e., species presence–absence data) and geographic distance depends on two parameters, the fundamental biodiversity number θ and the dispersal rate *m*. The shape of the curve also depends on the size of the sampling units (Hubbell 2001: Chapter 7).

Following Nekola and White (1999), Condit et al. (2002) and Tuomisto et al. (2003*c*) presented similarity decay plots to model the decrease of species composition similarities among sites as a function of geographic distances, a typical application of the distance-based approach. These plots are interpreted like correlograms. Models can be fitted to the data, corresponding to various hypotheses. The coefficient of determination (R^2) indicates the fit, or degree of adjustment of the data, to the model corresponding to the hypothesis. Tests of significance of this R^2 can be computed using either the Mantel test or statistical procedures that use Mantel-like permutations, such as the partial Mantel test or regression on distance matrices. Another interesting application is that of Tuomisto et al. (2003*a*), who wanted to relate reflectance differences, visible in satellite images, to differences in vegetation composition, irrespective of the specific vegetation, in order to draw rough maps delineating vegetation patches in the tropical forest using satellite data.

Tuomisto et al. (2003*c*) used Mantel and partial Mantel tests to compare competing hypotheses explaining the variation *among floristic distances*, which is appropriate. They also used multiple regression on distance matrices to estimate the relative contributions (R^2) of competing hypotheses to the variation of the floristic distances. These procedures are valid for testing hypotheses about the causes of the variation of the distances, such as the predictions of Hubbell's model, but perhaps not for parameter estimation. In any case, they should not be construed as explanations of the among-sites variation in species composition, SS(**Y**), which lays in the variance of the raw species data.

Unanswered problems remain about the use of multiple regression on distance matrices to address such questions, particularly the lack of independence among the distances. The matrix permutation procedure allows us to carry out a test of significance that has correct Type I error, but what is the effect of the lack of independence on parameter estimation? What does the coefficient of determination mean, besides a measure of adjustment of a model to the distances? Can we interpret it in terms of a proportion of explained variation, as in ordinary regression analysis? Statistical developments, as well as simulation studies, are needed to determine the validity and the limits of application of that method for analyzing spatial variation in beta diversity measures, for example among regions. This procedure will have to be thoroughly validated before we use it to draw conclusions about complex ecological problems and to design environmental policies from these conclusions. We will not speculate any further, one way or another, in the present paper.

SIMULATION STUDY: CANONICAL ANALYSIS VS. MANTEL TEST

We carried out simulations to empirically assess the power of two partitioning methods, canonical and on distance matrices (Mantel approach), for the interpretation of the variation in community composition among sites. This is a level-2 question, according to

the nomenclature established at the beginning of the section *Analyzing beta diversity*. However, we have shown that several authors studying such questions have used the Mantel approach, which should be reserved for studying the variation in beta diversity among groups of sites (level-3 question); hence the pertinence of the comparison in the simulations reported in the present section. Rather than analyzing actual data sets, we chose the simulation approach because the use of simulations allows us to compare the outcome of the analysis to ''the truth,'' which we know because we generated it. The analysis of real data sets is often limited in terms of the number of data sets available with all the necessary variables; it is also limited by the fact that one does not know whether the null hypothesis H_0 (there is no effect of the explanatory variables on the species data) or the alternative hypothesis H_1 (there is an effect) is true in any particular case. Monte Carlo simulations allow researchers to know the exact relationship between variables in the data. No doubt exists as to which hypothesis, H_0 or H_1, is true in each particular simulated data set (Milligan 1996).

Data generation

The first task was to generate data tables having the following properties. (1) The species data had to contain significant spatial patterns (different from random) of known properties. (2) Optionally, the species data had to be made dependent upon environmental variables having known properties. (3) The species data had to resemble species presence–absence or abundance data found in real ecosystems. (4) The environment component, optionally, also had a spatial structure of known properties.

To answer these requirements, we created a new version of a simulation program used previously to study the consequences of spatial structures for the design of ecological field surveys (Legendre et al. 2002) and field experiments (Legendre et al. 2004). The program is available in Supplement 1. Through the use of the conditional sequential Gaussian simulation method implemented in subroutine SGSIM of the geostatistical software library GSLIB (Deutsch and Journel 1992), this program allowed us to generate species data with autocorrelation of known intensity under a particular variogram model. The program also allowed an environmental variable to have an influence, called β (beta) by reference to a regression coefficient, of known intensity on some of the species. The species also contained random $\mathcal{N}(0,1)$ ''innovation'' at each site. The model implemented by the program for the value of a species j at site i was

$$S_{ij} = \beta_j E_{ik} + \mathrm{SA}_{Sij} + \varepsilon_{Sij} \quad (3)$$

where E_{ik} is the value of environmental variable k at site i; β_j transfers the effect of environmental variable k to species j; SA_{Sij} is the added value given by spatial autocorrelation to species j at site i; and ε_{Sij} is the ''innovation'' value for species j at site i (normal error). Two groups of species were generated and written to the species data files: a group of species always unrelated to the environment and another group possibly related to the environmental variables. Environmental variable k may contain any combination (controlled by the list of parameters of the computer run) of the following three elements: a deterministic structure, spatial autocorrelation, and normal ''innovation'' at each site simulated by random normal deviates $\mathcal{N}(0,1)$.

The preliminary species data generated by Eq. 3 were positive or negative real numbers, fairly symmetrically distributed. The following four steps were necessary to transform them into final species-like data. (1) Species vectors containing the influence of an environmental variable had means and variances different from the species unrelated to environmental variables. To give equal importance to all species at this step, each species vector was standardized. (2) To obtain a combination of species with different mean abundances, a normal deviate with mean 0 and standard deviation 1.5 was drawn at random for each species; all values S_{ij} generated for species j were multiplied by that constant. (3) The resulting values y_{ij} were exponentiated, obtaining $y'_{ij} = \exp(y_{ij})$. (4) Values y'_{ij} were rounded to the lower integer or to presence–absence data. The resulting data tables resembled species abundance data found in nature, with about 50% zeros and good variation among species means. An example is provided in Supplement 2.

The simulated surface was a square grid containing $100 \times 100 = 10\,000$ points. The units of the grid are pixels of arbitrary size. After generating species and environmental variable values over the whole grid, we sampled it using a square regular grid design with $10 \times 10 = 100$ points and spacing of 10 units between immediate neighbors. Assuming that the simulated surface represents 1 km^2, or 100 ha, this is twice the surface area of the well-known Pasoh Reserve and Barro Colorado Nature Monument tropical forest plots (50 ha); each site on the grid is 10×10 m in size (or is located within a portion of territory of that size) and the sites pertaining to the sample are 100 m apart, forming a regular grid.

In all simulations, the first five species were simulated without environmental control and with autocorrelation controlled by a spherical variogram model with range 15 (arbitrary grid units) in both directions of the plane. A second set of five species was generated with different relationships to the environmental variables, as specified in sections A–E of Table 1 and Appendix B, Table B1 (also used to designate the panels of Fig. 3) and described as follows (A–E).

Section A.—A first set of simulations was meant to assess the Type I error of each method. The second group of five species was generated with autocorrelation controlled by the same variogram (range = 15) as

the first set of species. Five environmental variables were created containing $\mathcal{N}(0,1)$ deviates, but they had no influence on the species because the transfer parameter β from environment to species was set to 0. These simulations will allow us to assess the frequency of Type I error when testing the influence of the environmental variables on the species data (test of fraction [a + b]). Because that influence was set to $\beta = 0$, a valid test of significance should have a rejection rate equal to or smaller than the significance level α of the tests.

Section B.—In the second set of simulations, an independently generated environmental variable was made to influence each of five species with a transfer parameter $\beta = 0.5$. The environmental variables were autocorrelated with range of 15 in the two geographic directions of the plane; they also contained $\mathcal{N}(0,1)$ deviates. The data generation parameters of the stochastic simulations were chosen in such a way that the rejection rates in RDA were close to 1.

Section C.—The third set of simulations was similar to that of section B, with a lower transfer parameter $\beta = 0.25$.

Section D.—In the fourth set of simulations, an independently generated environmental variable was made to influence each of five species with a transfer parameter $\beta = 0.5$. The environmental variables contained a spatial planar gradient (deterministic structure), with values increasing from the lower-left to the upper-right corner of the plane; they also contained $\mathcal{N}(0,1)$ deviates. Again, the data generation parameters were chosen in such a way that the rejection rates in RDA were close to 1.

Section E.—The fifth set of simulations was similar to that of section D, with a lower transfer parameter $\beta = 0.25$.

Situations 2–5 are meant to compare the power of the two methods of variation partitioning in situations corresponding to hypotheses 2 and 3 stated in the *Introduction*. Two series of 1000 data sets were analyzed for each of these five situations: the first series with species abundances and the second with presence–absence data.

Data analysis

Each species data set was analyzed using two methods of variation partitioning against a set of environmental variables, described in the data generation section, and a set of spatial variables created as follows. For canonical partitioning, we used (1) the *X* and *Y* geographic coordinates of the 100 points forming the sampling grid, (2) a third-order polynomial function of the *X* and *Y* geographic coordinates, and (3) a set of PCNM variables selected by a forward-selection procedure similar to that of the program CANOCO (ter Braak and Šmilauer 1998). This resulted in a selection of 4.5 PCNM variables, on average, per data set (range 0–15). For partitioning on distance matrices, we used (1) a geographic distance matrix $\mathbf{D}(XY)$ computed from the *X* and *Y* geographic coordinates, (2) a distance matrix $\mathbf{D}$(polyn.) obtained by computing Euclidean distances based on the third-order polynomial of the geographic coordinates described in the previous paragraph, and (3) a distance matrix $\ln(\mathbf{D}(XY))$ obtained by taking the natural logarithm of the values in $\mathbf{D}(XY)$. The detailed analysis of one of the sets of data tables used in the simulations is presented in Appendix B for illustration.

The first and second analyses are comparable for the two methods: the first one uses the *X* and *Y* coordinates, raw or in the form of distances, whereas the second one uses the third-order polynomial function of *X* and *Y*, raw or in the form of distances. The third analysis is possibly each method's most favorable analysis and does not have an equivalent in the other method. For raw data, PCNM analysis allows the modeling of spatial relationships at multiple scales and is thus likely to account for more variation than the two forms of trend-surface analysis, linear and polynomial. For distance data generated under hypothesis 2 of the *Introduction*, Hubbell's model predicts a linear relationship between spatially autocorrelated species data and the log of geographic distances (Hubbell 2001: Fig. 7.9).

For canonical variation partitioning, the quantitative and presence–absence species data were Hellinger-transformed (Legendre and Gallagher 2001) prior to partitioning. The environmental variables were used without transformation. For partitioning on distance matrices, a Hellinger dissimilarity matrix was computed from the quantitative and presence–absence species data; the Hellinger distance on presence–absence data is equal to

$$\sqrt{2} \times \sqrt{(1 - \text{Ochiai similarity coefficient})}.$$

The Euclidean distance function was used to compute the environmental distance matrix.

The Hellinger transformation or distance was used in all cases to insure that the results were comparable across analyses. Legendre and Gallagher (2001) have shown that in linear methods such as canonical redundancy analysis, the Hellinger distance (instead of the Euclidean distance) among sites is preserved if the data have been Hellinger-transformed prior to analysis. The Hellinger transformation consists in transforming the presence–absence or abundance data into relative values per site, by dividing each value by the site sum, then taking the square root of the resulting values. [Because of Whittaker's (1972) suggestion of using the mean of Jaccard coefficients as a measure of beta diversity, the partitioning analyses were repeated using $D = (1 - \text{Jaccard similarity coefficient})$. Results in terms of rejection rates and fractions of community composition variation were quite similar to those that we will report, where the Hellinger dissimilarity was used.]

In the simulation program, three canonical analyses were necessary to compute the fractions of variation [a + b + c], [a + b], [b + c], [a], [b], and [c] of Fig. 1 and test them (except [b]) for significance. However, when using standard canonical analysis software, e.g., CANOCO, five separate analyses are necessary to test all fractions of variation, as described in Borcard et al. (1992) and in Legendre and Legendre (1998: Section 13.5). On the Mantel side, three regressions on distance matrices were necessary: a simple regression of species on environmental distances provided an estimate and test of fraction [a + b], a simple regression of species on geographic distances procured an estimate and test of fraction [b + c], whereas a multiple regression of species vs. environmental and geographic distances produced an estimate and test of fraction [a + b + c]. It also provided tests of fractions [a] and [c] because tests of the partial regression coefficient are equivalent to tests of partial correlation coefficients as could be done by partial Mantel tests; estimates of fractions [a], [b] and [c] in the distance world were obtained from [a + b + c], [a + b] and [b + c] by subtraction. Fraction [b] contains variation explained by the environmental variables that is also spatially structured; this fraction cannot be tested for significance because it is only obtained by subtraction, and not by estimation of an explicit parameter in the regression model (Borcard et al. 1992, Legendre and Legendre 1998). Tests of significance involved 999 random permutations in all cases.

Fraction [d], which contains residuals, is not reported, although it was quite high in all analyses. It is easy to compute as (1 − [a + b + c]). These high [d] values are due to our choice of data generation parameters: had we generated species data more highly determined by the environmental and spatial variables, all fractions would have been significant using the two methods of analysis and it would not have been possible to determine which one had higher power. So we chose, instead, to generate species and environmental data with weak spatial structures and small correlations between the environmental and spatial data.

Results and discussion

The simulation results, reported in Table 1 and Fig. 3 (numerical results in Appendix C, Table C1), show the following:

1. Fig. 3 and Appendix C Table C1.—Under all simulated conditions, regression on distance matrices highly underestimated the portions of the species variation explained by the environmental variables alone (Table C1, columns [a + b]), the spatial relationships alone (columns [b + c]), or both sources of variation (columns [a + b + c]). Dutilleul et al. (2000) showed, analytically and by simulations, that in the case of normal error, the Mantel correlation estimates the square of the correlation between two variables. Our simulations used more complex and spatially autocorrelated error components; nevertheless, the fractions of vari-

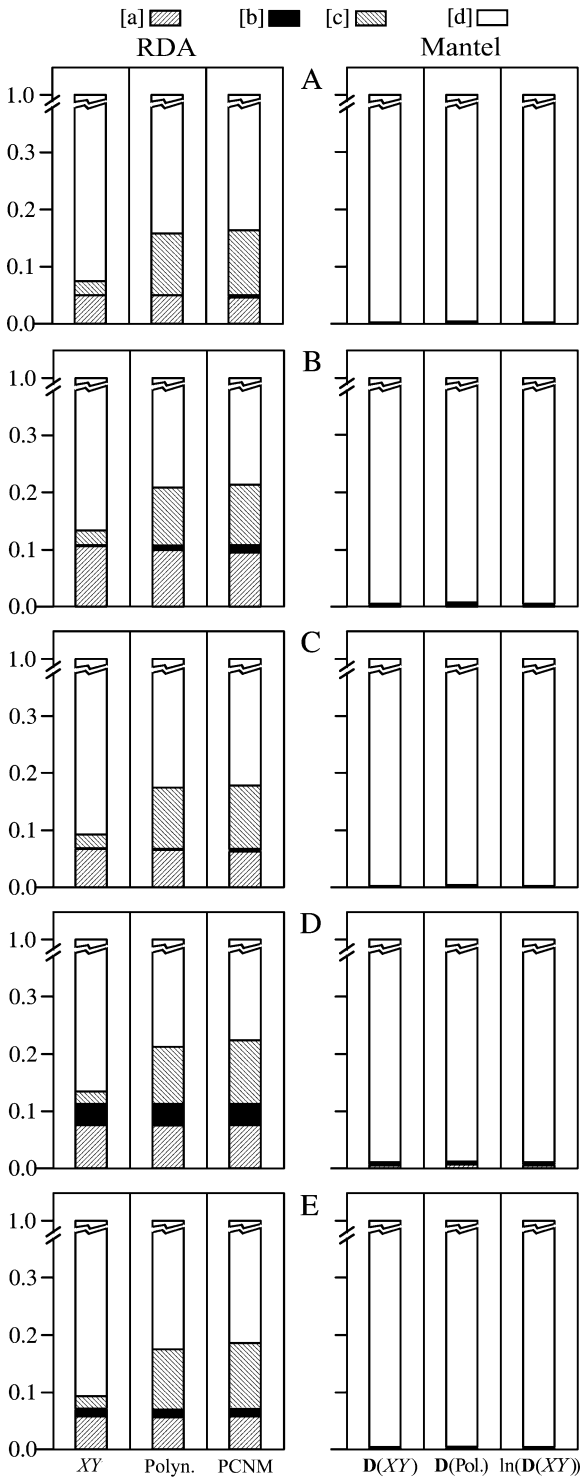

FIG. 3. Simulation results for presence–absence data: proportion of the species variation in the fractions shown in Fig. 1 (means after 1000 simulations). (A)–(E) are the different types of simulations. The *x*-axis shows different treatments of the geographic information. See *Simulation study: canonical analysis vs. Mantel test; Data analysis.* Detailed numerical results are presented in Appendix C, Table C1.

TABLE 1. Simulation results: rates of rejection of the null hypothesis, H_0 (trace of the fraction is 0), at the $\alpha = 5\%$ level after 1000 simulations.

Partitioning method	Species presence–absence					Species abundance				
	[a + b + c]	[a + b]	[b + c]	[a]	[c]	[a + b + c]	[a + b]	[b + c]	[a]	[c]
A) *S* unrelated to *E* ($\beta = 0$), *S* autocorrelated (range = 15), $E \sim \mathcal{N}(0,1)$; [a + b], Type I error; [b + c], power										
RDA, *XY*	0.109	0.040	0.214	0.042	0.203	0.102	0.045	0.191	0.045	0.168
RDA, polynomial	0.280	0.037	0.391	0.052	0.361	0.261	0.047	0.344	0.045	0.324
RDA, PCNM	0.915	0.037	0.989	0.039	0.983	0.916	0.056	0.992	0.059	0.987
Mantel, **D**(*XY*)	0.067	0.041	0.115	0.040	0.115	0.106	0.067	0.133	0.066	0.133
Mantel, **D**(polyn.)	0.062	0.047	0.071	0.046	0.071	0.068	0.064	0.082	0.063	0.080
Mantel, ln(**D**(*XY*))	0.073	0.039	0.171	0.039	0.170	0.134	0.066	0.222	0.066	0.225
B) *S* related to *E* ($\beta = 0.5$), *S* autocorrelated (range = 15), *E* autocorrelated (range = 15)										
RDA, *XY*	0.997	0.997	0.211	0.997	0.199	0.963	0.973	0.194	0.964	0.181
RDA, polynomial	0.989	0.997	0.358	0.996	0.314	0.944	0.971	0.327	0.950	0.304
RDA, PCNM	1.000	0.997	0.995	0.995	0.979	0.999	0.970	0.993	0.954	0.986
Mantel, **D**(*XY*)	0.296	0.281	0.104	0.281	0.107	0.501	0.487	0.124	0.478	0.121
Mantel, **D**(polyn.)	0.213	0.285	0.088	0.288	0.083	0.386	0.492	0.081	0.488	0.081
Mantel, ln(**D**(*XY*))	0.308	0.279	0.161	0.281	0.153	0.529	0.488	0.199	0.481	0.188
C) *S* related to *E* ($\beta = 0.25$), *S* autocorrelated (range = 15), *E* autocorrelated (range = 15)										
RDA, *XY*	0.536	0.510	0.188	0.498	0.192	0.490	0.426	0.187	0.429	0.182
RDA, polynomial	0.615	0.505	0.370	0.428	0.328	0.559	0.432	0.342	0.380	0.313
RDA, PCNM	0.982	0.514	0.993	0.414	0.985	0.981	0.434	0.993	0.366	0.987
Mantel, **D**(*XY*)	0.099	0.080	0.089	0.082	0.091	0.141	0.127	0.121	0.123	0.122
Mantel, **D**(polyn.)	0.082	0.074	0.078	0.073	0.077	0.108	0.122	0.083	0.121	0.084
Mantel, ln(**D**(*XY*))	0.102	0.077	0.148	0.075	0.149	0.179	0.119	0.190	0.118	0.186
D) *S* related to *E* ($\beta = 0.5$), *S* autocorrelated (range = 15), *E* with deterministic gradient										
RDA, *XY*	0.991	0.995	0.974	0.785	0.118	0.955	0.971	0.922	0.733	0.091
RDA, polynomial	0.987	0.996	0.903	0.750	0.291	0.951	0.970	0.848	0.683	0.277
RDA, PCNM	1.000	0.998	0.999	0.845	0.962	1.000	0.969	0.999	0.794	0.959
Mantel, **D**(*XY*)	0.593	0.586	0.578	0.357	0.175	0.749	0.741	0.698	0.570	0.222
Mantel, **D**(polyn.)	0.430	0.585	0.140	0.501	0.066	0.651	0.740	0.223	0.675	0.068
Mantel, ln(**D**(*XY*))	0.638	0.582	0.741	0.370	0.293	0.776	0.739	0.798	0.574	0.366
E) *S* related to *E* ($\beta = 0.25$), *S* autocorrelated (range = 15), *E* with deterministic gradient										
RDA, *XY*	0.517	0.534	0.563	0.206	0.111	0.454	0.486	0.513	0.202	0.083
RDA, polynomial	0.631	0.539	0.608	0.196	0.294	0.554	0.481	0.538	0.189	0.270
RDA, PCNM	0.980	0.536	0.997	0.294	0.982	0.979	0.492	0.996	0.246	0.979
Mantel, **D**(*XY*)	0.149	0.147	0.204	0.094	0.111	0.256	0.237	0.279	0.135	0.132
Mantel, **D**(polyn.)	0.118	0.149	0.091	0.119	0.065	0.193	0.233	0.115	0.189	0.075
Mantel, ln(**D**(*XY*))	0.172	0.148	0.322	0.090	0.194	0.296	0.232	0.400	0.129	0.217

Notes: Abbreviations are: *S*, species; *E*, environmental variables; *XY*, geographic coordinates of the sites. Range is the range parameter of the variogram for generation of autocorrelation; the size of the simulation field was 100×100 (arbitrary units).

ation reported in Table C1 are in general agreement with the findings of Dutilleul et al. The coefficients of determination (R^2) obtained by regression on distance matrices are thus clearly not estimates of the proportions of variation explained by the explanatory variables.

Appendix A shows that the portions of variation estimated by canonical analysis are statistically correct estimates of fractions of the community composition variation (beta diversity). Hence, regression on distance matrices, applied to data simulated to be at least partly similar to the data collected to study the origin of variation in species composition among sites, produced incorrect partitioning of that variation.

The objective of these simulations was to compare RDA to regression on distance matrices, not RDA using a cubic polynomial function of the geographic coordinates to PCNM analysis. PCNM analysis always produced a portion of explained variation not much larger than RDA using the polynomial function. Forward selection was applied in the case of PCNM analysis, and resulted in the selection of 4.5 PCNM variables on average, whereas RDA using the polynomial was always conducted with the nine monomials. The adjusted coefficient of determination, R^2_{adj}, is more suitable than R^2 for comparing regression or canonical analysis results obtained using different numbers of objects and explanatory variables. Thus, R^2_{adj} would provide a fairer representation of the differences in explained variation between sets of simulations. For the first group of simulations, for example (A in Fig. 3 and Table C1), R^2_{adj} for the total explained variation [a + b + c] is 0.0053 for RDA using *XY*, 0.0200 for RDA using the polynomial, and 0.0742 for RDA using PCNM after forward selection. In all of our simulations, PCNM analysis always extracted more information from the species data than RDA based upon a cubic polynomial of the geographic coordinates. R^2_{adj} could not be applied to the

whole simulation study, however, because we do not know how to calculate it in the case of the Mantel test.

It is interesting to note that there was not much difference between the partitioning results obtained for the simulated presence–absence and abundance data.

2. Table 1A, columns [a + b] and [a].—Type I error, which is the frequency of rejection of the null hypothesis when H_0 is true, was correct for both methods. The significance level used in the simulation tests, $\alpha = 0.05$, was inside the 95% confidence intervals of the rejection rates in all cases. Had the species and environmental data sets both been spatially structured, we would have expected the tests to be too liberal and the rejection rates to be higher than α (property shown for partial Mantel test by Oden and Sokal 1992, and for correlation and regression analysis by Legendre et al. 2002).

3. Table 1.—In all simulations in which an effect was present (Table 1A, columns [b + c] and [c], and Table 1B–E), canonical analysis rejected the null hypothesis with frequencies much higher than regression on distance matrices. That was the case, in particular, when comparing each method's most favorable form of spatial analysis, i.e., PCNM analysis in the canonical case and regression on logged distances in the Mantel case. Canonical analysis clearly has much higher power than regression on distance matrices to detect effects of environmental or spatial explanatory variables on species data.

In Table C1(A), one may wonder why the environmental variables (fractions [a + b]) explain more of the species variation in RDA than in Mantel tests; in this set of simulations, there was no effect of the environmental variables on the species ($\beta = 0$). The reason is that five environmental variables were present in the analysis in RDA, whereas in regression on distance matrices these variables were combined into a single distance matrix. A random variable in regression explains, on average, a portion $1/(n - 1)$ of the response data by chance alone. Thus, using five environmental variables, we expect RDA to explain by chance, on average, $5/(100 - 1) = 0.0505$ of the species data in our simulations. This is very close to the values reported in the table. By this criterion, the variation explained by linear trend surface analysis (RDA on the *X* and *Y* geographic coordinates) in columns [b + c] is random in sections a, b, and c of the table because the reported values are close to 0.0202, the value expected from regression on two random variables. In sections d and e of the table, the linear trend surface (RDA, *XY*) has slightly more success, producing explained variation from 0.03 to 0.06 in columns [b + c]. This is because the environmental variables contained a deterministic gradient in these simulations.

In regression on distance matrices, a single variable would explain by chance, on average, $1/(100 - 1) = 0.0101$ of the species data, assuming that a random effect turned into a distance matrix behaves in the same way in regression on distance matrices as in ordinary multiple regression or RDA. This is slightly higher than the observed values. Accounting for the effect reported by Dutilleul et al. (2000) and described in paragraph 1 of this subsection, the portion of explained variation should be $0.0101^2 = 0.0001$, which is slightly lower than the values reported in the table.

Conclusion

In recent years, several authors who were studying the variation in community composition among sites (beta diversity) have used variation partitioning on distance matrices (Mantel approach) instead of canonical partitioning. These authors have certainly used partitioning on distance matrices in good faith, after discovering that partial Mantel tests (Smouse et al. 1986) or our program for regression on distance matrices (Casgrain 2001) would do the job. Their results triggered our interest and led us to compare variation partitioning done in two different ways, on raw data and on distance matrices, for data corresponding to questions concerning the variation in community composition among sites (level-2 questions in *Analyzing beta diversity*), generated under hypotheses about the origin of the variation of community composition among sites. The generated data correspond, at least in part, to hypotheses 2 and 3 of the *Introduction* about the origin of beta diversity.

The theoretical developments and simulation results reported in this paper allow us to conclude the following. (1) The variance of the community composition table is a measure of beta diversity. (2) The variance of a dissimilarity matrix among sites is not the variance of the community composition table, or a measure of beta diversity; hence, partitioning on distance matrices should not be used to study the variation in community composition among sites. (3) In all of our simulations, partitioning on distance matrices underestimated the amount of variation in community composition explained by the raw data approach. (4) The tests of significance had less power than the tests of canonical ordination. Hence, the proper statistical procedure for partitioning the spatial variation of community composition data among environmental and spatial components, and for testing hypotheses about the origin and maintenance of variation in community composition among sites, is canonical partitioning. The Mantel approach is appropriate for testing hypotheses concerning the variation in beta diversity among groups of sites (level-3 hypotheses). Regression on distance matrices is also appropriate to fit models to similarity decay plots. It should not be used for questions related to the variation in the raw community composition data among sites.

Our conclusions about the domain of application of the Mantel test apply to the analysis of similarity (ANOSIM) as well, a test of significance of matrix correlation used in marine biology. ANOSIM (Clarke 1988,

1993) is simply a form of the Mantel test (Legendre and Legendre 1998: section 10.5).

The simulations reported in this paper provide important results regarding the Type I error and power of the permutation test commonly used in canonical analysis when applied to community composition variation. We also tested the power of PCNMs in spatial analysis and concluded that they provide a much better representation of spatial structures than other procedures commonly used by ecologists, such as trend surface analysis.

Simulated data are always a simplification compared to the complexity of ecological data observed in the field. This is certainly true for the data analyzed in this paper, despite our efforts to build into them the key properties necessary for the purpose of comparing the two variation partitioning procedures: species data structure close to those observed in nature, different intensities of the relationships between species and environmental variables (including the absence of such relationships), spatial structure due to autocorrelation in the species data, and to autocorrelation or deterministic processes in the environmental data.

We hope that this paper will reach the ecologists who will be producing the environmental studies on which world deciders may base scientific management decisions about biodiversity in endangered environments.

Acknowledgments

This paper has greatly benefited from discussions and detailed comments by Hanna Tuomisto and Kalle Ruokolainen, and from interesting comments from two anonymous reviewers. Access to the computers of the *Environnement Scientifique Intégré* (ESI) of Université de Montréal for our simulation work is gratefully acknowledged, with special thanks to Bernard Lorazo. This research was supported by NSERC grant number OGP0007738 to P. Legendre.

Literature Cited

Allan, J. D. 1975. Components of diversity. Oecologia **18**: 359–367.

Anderson, M. J., and N. A. Gribble. 1998. Partitioning the variation among spatial, temporal and environmental components in a multivariate data set. Australian Journal of Ecology **23**:158–167.

Bell, G. 2001. Neutral macroecology. Science **293**:2413–2418.

Borcard, D., and P. Legendre. 1994. Environmental control and spatial structure in ecological communities: an example using oribatid mites (Acari, Oribatei). Environmental and Ecological Statistics **1**:37–53.

Borcard, D., and P. Legendre. 2002. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. Ecological Modelling **153**:51–68.

Borcard, D., P. Legendre, C. Avois-Jacquet, and H. Tuomisto. 2004. Dissecting the spatial structure of ecological data at multiple scales. Ecology **85**:1826–1832.

Borcard, D., P. Legendre, and P. Drapeau. 1992. Partialling out the spatial component of ecological variation. Ecology **73**:1045–1055.

Bray, R. J., and J. T. Curtis. 1957. An ordination of the upland forest communities of southern Wisconsin. Ecological Monographs **27**:325–349.

Casgrain, P. 2001. Permute! Version 3.4 user's manual. Département de sciences biologiques, Université de Montréal, Montreal, Quebec, Canada.

Clarke, K. R. 1988. Detecting change in benthic community structure. Pages 131–142 *in* R. Oger, editor. Proceedings of Invited Papers, 14th International Biometric Conference, Namur, Belgium. Société Adolphe Quételet, Gembloux, Belgium.

Clarke, K. R. 1993. Non-parametric multivariate analyses of changes in community structure. Australian Journal of Ecology **18**:117–143.

Condit, R., N. et al. 2002. Beta-diversity in tropical forest trees. Science **295**:666–669.

Decaens, T., and J. P. Rossi. 2001. Spatio-temporal structure of earthworm community and soil heterogeneity in a tropical pasture. Ecography **24**:671–682.

Deutsch, C. V., and A. G. Journel. 1992. GSLIB: geostatistical software library and user's guide. Oxford University Press, New York, New York, USA.

Dutilleul, P., J. D. Stockwell, D. Frigon, and P. Legendre. 2000. The Mantel-Pearson paradox: statistical considerations and ecological implications. Journal of Agricultural, Biological and Environmental Statistics **5**(2):131–150.

Erkkilä, H. M. J. B. 1998. Seed banks of grazed and ungrazed Baltic seashore meadows. Journal of Vegetation Science **9**: 395–408.

Fortin, M.-J., and J. Gurevitch. 1993. Mantel tests: spatial structure in field experiments. Pages 342–359 *in* S. M. Scheiner and J. Gurevitch, editors. Design and analysis of ecological experiments. Chapman and Hall, New York, New York, USA.

Gauch, H. G., Jr., and R. H. Whittaker. 1972. Coenocline simulation. Ecology **53**:446–451.

Gentry, A. H. 1988. Changes in plant community diversity and floristic composition on environmental and geographical gradients. Annals of the Missouri Botanical Garden **75**: 1–34.

Greenberg, J. H. 1956. The measurement of linguistic diversity. Language **32**:109–115.

He, F. 2005. Deriving a neutral model of species abundance from fundamental mechanisms of population dynamics. Functional Ecology **19**:187–193.

Hill, M. O. 1973. Diversity and evenness: a unifying notation and its consequences. Ecology **54**:427–432.

Hubbell, S. P. 2001. The unified neutral theory of biodiversity and biogeography. Princeton University Press, Princeton, New Jersey, USA.

Hutchinson, G. E. 1957. Concluding remarks. Cold Spring Harbor Symposia on Quantitative Biology **22**:15–427.

Jutila, H. M. 2002. Seed banks of river delta meadows on the west coast of Finland. Annales Botanici Fennici **39**: 49–61.

Kerlinger, F. N., and E. J. Pedhazur. 1973. Multiple regression in behavioral research. Holt, Rinehart and Winston, New York, New York, USA.

Kiflawi, M., and M. Spencer. 2004. Confidence intervals and hypothesis testing for beta diversity. Ecology **85**:2895–2900.

Koch, L. F. 1957. Index of biotal dispersity. Ecology **38**:145–148.

Lande, R. 1996. Statistics and partitioning of species diversity, and similarity among multiple communities. Oikos **76**: 5–13.

Legendre, P. 1990. Quantitative methods and biogeographic analysis. Pages 9–34 *in* D. J. Garbary and R. G. South, editors. Evolutionary biogeography of the marine algae of the North Atlantic. NATO ASI Series. Volume G 22. Springer-Verlag, Berlin, Germany.

Legendre, P. 1993. Spatial autocorrelation: trouble or new paradigm? Ecology **74**:1659–1673.

Legendre, P. 2000. Comparison of permutation methods for the partial correlation and partial Mantel tests. Journal of Statistical Computation and Simulation **67**:37–73.
Legendre, P., and M. J. Anderson. 1999. Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. Ecological Monographs **69**:1–24.
Legendre, P., M. R. T. Dale, M.-J. Fortin, P. Casgrain, and J. Gurevitch. 2004. Effects of spatial structures on the results of field experiments. Ecology **85**:3202–3214.
Legendre, P., M. R. T. Dale, M.-J. Fortin, J. Gurevitch, M. Hohn, and D. Myers. 2002. The consequences of spatial structure for the design and analysis of ecological field surveys. Ecography **25**:601–615.
Legendre, P., and E. D. Gallagher. 2001. Ecologically meaningful transformations for ordination of species data. Oecologia **129**:271–280.
Legendre, P., F.-J. Lapointe, and P. Casgrain. 1994. Modeling brain evolution from behavior: a permutational regression approach. Evolution **48**:1487–1499.
Legendre, P., and L. Legendre. 1998. Numerical ecology. Second English edition. Elsevier, Amsterdam, The Netherlands.
Legendre, P., and M. Troussellier. 1988. Aquatic heterotrophic bacteria: modeling in the presence of spatial autocorrelation. Limnolology and Oceanography **33**:1055–1067.
Levins, R. 1968. Evolution in changing environments: some theoretical explorations. Princeton University Press, Princeton, New Jersey, USA.
MacArthur, R., H. Recher, and M. Cody. 1966. On the relation between habitat selection and species diversity. American Naturalist **100**:319–332.
Magnan, P., M. A. Rodríguez, P. Legendre, and S. Lacasse. 1994. Dietary variation in a freshwater fish species: relative contribution of biotic interactions, abiotic factors, and spatial structure. Canadian Journal of Fisheries and Aquatic Sciences **51**:2856–2865.
Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Research **27**:209–220.
Méot, A., P. Legendre, and D. Borcard. 1998. Partialling out the spatial component of ecological variation: questions and propositions in the linear modeling framework. Environmental and Ecological Statistics **5**:1–27.
Milligan, G. W. 1996. Clustering validation—results and implications for applied analyses. Pages 341–375 *in* P. Arabie, L. J. Hubert, and G. De Soete, editors. Clustering and classification. World Scientific Publishing, River Edge, New Jersey, USA.
Mulder, C., D. De Zwart, H. J. Van Wijnen, A. J. Schouten, and A. M. Breure. 2003. Observational and simulated evidence of ecological shifts within the soil nematode community of agroecosystems under conventional and organic farming. Functional Ecology **17**:516–525.
Nekola, J. C., and P. S. White. 1999. The distance decay of similarity in biogeography and ecology. Journal of Biogeography **26**:867–878.
Oden, N. L., and R. R. Sokal. 1992. An investigation of three-matrix permutation tests. Journal of Classification **9**:275–290.
Odum, E. P. 1950. Bird populations of the Highlands (North Carolina) plateau in relation to plant succession and avian invasion. Ecology **31**:587–605.
Økland, R. H. 2003. Partitioning the variation in a plot-by-species data matrix that is related to *n* sets of explanatory variables. Journal of Vegetation Science **14**:693–700.
Orwig, D. A., D. R. Fosterand, and D. L. Mausel. 2002. Landscape patterns of hemlock decline in New England due to the introduced hemlock woolly adelgid. Journal of Biogeography **29**:1475–1487.
Parris, K. M., and M. A. McCarthy. 1999. What influences the structure of frog assemblages at forest streams? Australian Journal of Ecology **24**:495–502.
Paszkowski, C. A., and W. M. Tonn. 2000. Community concordance between the fish and aquatic birds of lakes in northern Alberta, Canada: the relative importance of environmental and biotic factors. Freshwater Biology **43**: 421–437.
Pélissier, R., P. Couteron, S. Dray, and D. Sabatier. 2003. Consistency between ordination techniques and diversity measurements: two strategies for species occurrence data. Ecology **84**:242–251.
Phillips, O. L., P. N. Vargas, A. L. Monteagudo, A. P. Cruz, M. E. C. Zans, W. G. Sanchez, M. Yli-Halla, and S. Rose. 2003. Habitat association among Amazonian tree species: a landscape-scale approach. Journal of Ecology **91**:757–775.
Pitman, N. C. A., J. Terborgh, M. R. Silman, and P. Nuñez V. 1999. Tree species distributions in an upper Amazonian forest. Ecology **80**:2651–2661.
Pitman, N. C. A., J. W. Terborgh, M. R. Silman, P. Núñez V., D. A. Neill, C. E. Cerón, W. A. Palacios, and M. Aulestia. 2001. Dominance and distribution of tree species in two upper Amazonian terra firme forests. Ecology **82**:2101–2117.
Potts, M. D., P. S. Ashton, L. S. Kaufman, and J. B. Plotkin. 2002. Habitat patterns in tropical rain forests: a comparison of 105 plots in Northwest Borneo. Ecology **83**:2782–2797.
Pozzi, S., and D. Borcard. 2001. Effects of dry grassland management on spider (Arachnida: Araneae) communities on the Swiss occidental plateau. Écoscience **8**:32–44.
Pusey, B. J., M. J. Kennard, and A. H. Arthington. 2000. Discharge variability and the development of predictive models relating stream fish assemblage structure to habitat in northeastern Australia. Ecology of Freshwater Fish **9**: 30–50.
Quinghong, L., and S. Bråkenhielm. 1995. A statistical approach to decompose ecological variation. Water Air and Soil Pollution **85**:1587–1592.
Rao, C. R. 1964. The use and interpretation of principal component analysis in applied research. Sankhyaá, Series A **26**:329–358.
Reynolds, C. E., and G. Houle. 2003. Mantel and partial Mantel tests suggest some factors that may control the local distribution of *Aster laurentianus* at Iles de la Madeleine, Quebec. Plant Ecology **164**:19–27.
Shannon, C. E. 1948. A mathematical theory of communications. Bell System Technical Journal **27**:379–423.
Simpson, E. H. 1949. Measurement of diversity. Nature (London) **163**:688.
Smouse, P. E., J. C. Long, and R. R. Sokal. 1986. Multiple regression and correlation extensions of the Mantel test of matrix correspondence. Systematic Zoology **35**:627–632.
Somerfield, P. J., and J. D. Gage. 2000. Community structure of the benthos in Scottish Sea-lochs. IV. Multivariate spatial pattern. Marine Biology **136**:1133–1145.
Spencer, M., S. S. Schwartz, and L. Blaustein. 2002. Are there fine-scale spatial patterns in community similarity among temporary freshwater pools? Global Ecology and Biogeography **11**:71–78.
Su, J. C., D. M. Debinski, M. E. Jakubauskas, and K. Kindscher. 2004. Beyond species richness: community similarity as a measure of cross-taxon congruence for coarse-filter conservation. Conservation Biology **18**:167–173.
Svenning, J. C. 1999. Microhabitat specialization in a species-rich palm community in Amazonian Ecuador. Journal of Ecology **87**:55–65.
ter Braak, C. J. F. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology **67**:1167–1179.

ter Braak, C. J. F. 1987*a*. The analysis of vegetation—environment relationships by canonical correspondence analysis. Vegetatio **69**:69–77.
ter Braak, C. J. F. 1987*b*. Calibration. Pages 78–90 *in* R. H. G. Jongman, C. J. F. ter Braak, and O. F. R. van Tongeren, editors. Data analysis in community and landscape ecology. Pudoc, Wageningen, The Netherlands. [Reissued in 1995 by Cambridge University Press, Cambridge, UK.]
ter Braak, C. J. F., and A. P. Schaffers. 2004. Co-correspondence analysis: a new ordination method to relate two community compositions. Ecology **85**:834–846.
ter Braak, C. J. F., and P. Šmilauer. 1998. CANOCO reference manual and user's guide to CANOCO for Windows. Software for canonical community ordination. Version 4. Centre for Biometry, Wageningen, The Netherlands, and Microcomputer Power, Ithaca, New York, USA.
Tuomisto, H., A. D. Poulsen, K. Ruokolainen, R. C. Moran, C. Quintana, J. Celi, and G. Canas. 2003*a*. Linking floristic patterns with soil heterogeneity and satellite imagery in Ecuadorian Amazonia. Ecological Applications **13**:352–371.
Tuomisto, H., K. Ruokolainen, M. Aguilar, and A. Sarmiento. 2003*b*. Floristic patterns along a 43-km long transect in an Amazonian rain forest. Journal of Ecology **91**:743–756.
Tuomisto, H., K. Ruokolainen, R. Kalliola, A. Linna, W. Danjoy, and Z. Rodriguez. 1995. Dissecting Amazonian biodiversity. Science **269**:63–66.
Tuomisto, H., K. Ruokolainen, and M. Yli-Halla. 2003*c*. Dispersal, environment, and floristic variation of western Amazonian forests. Science **299**:241–244.
Veech, J. A., K. S. Summerville, T. O. Crist, and J. C. Gering. 2002. The additive partitioning of species diversity: recent revival of an old idea. Oikos **99**:3–9.
Whittaker, R. H. 1952. A study of summer foliage insect communities in the Great Smoky Mountains. Ecological Monographs **22**:1–44.
Whittaker, R. H. 1956. Vegetation of the Great Smoky Mountains. Ecological Monographs **26**:1–80.
Whittaker, R. H. 1960. Vegetation of the Siskiyou Mountains, Oregon and California. Ecological Monographs **30**:279–338.
Whittaker, R. H. 1972. Evolution and measurement of species diversity. Taxon **21**:213–251.
Williams, L. R., C. M. Taylor, M. L. Warren, and J. A. Clingenpeel. 2003. Environmental variability, historical contingency, and the structure of regional fish and macroinvertebrate faunas in Ouachita Mountain stream systems. Environmental Biology of Fishes **67**:203–216.

APPENDIX A

Canonical variation partitioning with statistical details showing that canonical variation partitioning, as per Borcard et al. (1992), provides a correct partitioning of the variation of a response data table **Y**, is available in ESA's Electronic Data Archive: *Ecological Archives* M075-017-A1.

APPENDIX B

Partitioning examples with an explanation, a table, and three figures showing the analysis of one of the sets of data tables used in the simulation study, are available in ESA's Electronic Data Archive: *Ecological Archives* M075-017-A2.

APPENDIX C

A table showing simulation results is available in ESA's Electronic Data Archive: *Ecological Archives* M075-017-A3.

SUPPLEMENT 1

Data generation and analysis programs used in the simulation study are available in ESA's Electronic Data Archive: *Ecological Archives* M075-017-S1.

SUPPLEMENT 2

Data tables used in the example detailed in Appendix B are available in ESA's Electronic Data Archive: *Ecological Archives* M075-017-S2.

Appendices to:

Legendre, P., D. Borcard, and P. Peres-Neto. 2005. Analyzing beta diversity: partitioning the spatial variation of community composition data. *Ecological Monographs* 75: 435-450.

APPENDIX A

Ecological Archives M075-017-A1

CANONICAL VARIATION PARTITIONING: STATISTICAL DETAILS

This Appendix shows that canonical variation partitioning, as per Borcard et al. (1992), provides a correct partitioning of the variation of a response data table **Y**.

The variation in a single response variable **y** is measured by the sum of the squared differences of the values to the mean of that variable (SS). If **y** was centered before the calculations, $SS(\mathbf{y}) = \sum y_i^2$. If we analyze **y** by ordinary least-squares regression on an explanatory variable **x**, also centered, the fitted values can be computed as $\hat{\mathbf{y}} = \mathbf{x}[\mathbf{x}'\mathbf{x}]^{-1}\mathbf{x}'\mathbf{y}$. The amount of variation in the fitted values, $SS(\hat{\mathbf{y}})$, can be computed in the same way as for $SS(\mathbf{y})$. The portion of the variation of **y** explained by **x** is given by the coefficient of determination, $R^2 = SS(\hat{\mathbf{y}}) / SS(\mathbf{y})$.

Let us move to multivariate data. Canonical redundancy analysis (RDA) of a response table **Y** by an explanatory table **X** consists in two steps: (1) a series of multiple regressions of the individual variables of **Y** on **X**, which produces a table of fitted values $\hat{\mathbf{Y}}$; this is followed by (2) a principal component analysis of $\hat{\mathbf{Y}}$ which produces the canonical eigenvalues and eigenvectors and the canonical ordination scores (Legendre and Legendre 1998, Section 11.1). The second step is not necessary for variation partitioning. Assuming that the individual variables in **Y** were centered on their means, the total variation in **Y**, $SS(\mathbf{Y})$, is obtained by computing the sum of the squared values in table **Y**, just as in the univariate case. Because **Y** was centered, the individual columns of $\hat{\mathbf{Y}}$ are also centered on zero; $SS(\hat{\mathbf{Y}})$ is obtained in the same way as $SS(\mathbf{Y})$, by computing the sum of the squared values in $\hat{\mathbf{Y}}$. The portion of the variation of **Y** explained by **X**, $SS(\hat{\mathbf{Y}}) / SS(\mathbf{Y})$, is the bimultivariate redundancy statistic $R^2_{\mathbf{Y}\hat{\mathbf{o}}\mathbf{X}}$ (Miller and Farr 1971), which is the canonical equivalent of the coefficient of determination R^2; it is called the RDA "trace" statistic in the program Canoco (ter Braak and Smilauer 2002). Canonical correspondence analysis (CCA) is similar to RDA, with two small differences: (1) the table subjected to analysis is not **Y** but a table, called $\overline{\mathbf{Q}}$ by Legendre and Legendre (1998, Section 11.2), which contains a transformation of the original species presence-absence or abundance data into contributions to chi-square; and (2) the regression involves weights, given by the row sums of table **Y** divided by the sum total of the values in **Y**, and written in a diagonal matrix of weights which intervenes in the regression equations. These weights are also taken into account when standardizing the explanatory variables **X** at the beginning of the analysis. The rest of the calculations are similar to RDA. This short exposé shows that

canonical analysis produces estimates of the portion of the variation of $\mathbf{Y}$ explained by $\mathbf{X}$ that are similar to the familiar coefficients of determination of regression analysis.

All the fractions of variation that will be displayed in the simulation result tables (see the "Simulation study" section of the main paper) were obtained from 3 multiple regressions (for a single **y** response variable) or 3 canonical analyses (when analyzing a multivariate response table **Y**), followed by simple calculations: (1) RDA($\mathbf{Y}$|environmental matrix $\mathbf{X}_1$) produces the bimultivariate R^2 (or "trace") statistic $\mathrm{SS}(\hat{\mathbf{Y}}) / \mathrm{SS}(\mathbf{Y})$ for [a+b], (2) RDA($\mathbf{Y}$|spatial matrix $\mathbf{X}_2$) produces the statistic $\mathrm{SS}(\hat{\mathbf{Y}}) / \mathrm{SS}(\mathbf{Y})$ for [b+c], and (3) RDA($\mathbf{Y}$|matrices $\mathbf{X}_1$ and $\mathbf{X}_2$) produces the statistic $\mathrm{SS}(\hat{\mathbf{Y}}) / \mathrm{SS}(\mathbf{Y})$ for [a+b+c]. From these results, one can calculate [a] = [a+b+c] – [b+c], [c] = [a+b+c] – [a+b], and [b] = [a+b] + [b+c] – [a+b+c] (Fig. 1 of the main paper). The residual variation is given by the bimultivariate coefficient of nondetermination, [d] = 1 – [a+b+c]. The fractions [a], [b], [c], and [d] are additive and sum to 1, as in partial regression analysis (see for instance Legendre and Legendre 1998, Subsection 10.3.5).

While calculating the fractions of variation only involves simple canonical analyses, partial canonical analyses are necessary to test the significance of fractions [a] and [c] of variation partitioning. Partial canonical analysis is the direct extension of partial regression to multivariate response data.

LITERATURE CITED

Borcard, D., P. Legendre, and P. Drapeau. 1992. Partialling out the spatial component of ecological variation. Ecology 73: 1045-1055.

Legendre, P., and L. Legendre. 1998. Numerical ecology, 2nd English edition. Elsevier, Amsterdam, The Netherlands.

Miller, J. K., and S. D. Farr. 1971. Bimultivariate redundancy: a comprehensive measure of interbattery relationship. Multivariate Behavioral Research 6: 313-324.

ter Braak, C. J. F. and P. Smilauer. 2002. Canoco reference manual and CanoDraw for Windows user's guide: software for canonical community ordination (version 4.5). Microcomputer Power, Ithaca, New York.

APPENDIX B

Ecological Archives M075-017-A2

PARTITIONING EXAMPLES:
AN EXPLANATION, A TABLE (TABLE B1) AND THREE FIGURES (FIG. B1, FIG. B2, FIG. B3)

This Appendix contains the analysis of one of the sets of data tables used in the simulation study. Readers will be able to examine a series of generated data tables, and appreciate the steps of the statistical analysis.

Two tables of species and one table of environmental data were generated using the stochastic procedure described in the "Data generation" subsection. The data sets contained 100 rows corresponding to 100 sites forming a 10 x 10 regular grid in the 100 x 100 field. Five environmental variables were generated, each one with a deterministic structure (gradient), autocorrelation controlled by a spherical variogram model with a range of 15, and standard normal deviates $N(0,1)$ at each site. This example thus does not correspond to the simulations, where the environmental variables had either spatial autocorrelation or a deterministic structure, not both. The random normal deviates contributed 472 units to the sum of squares, the spatial autocorrelation component 482 units, and the deterministic slope 537 units, for a total sum of squares of 1491 in the environmental variables. These variables were normally distributed. Five species were created with random normal deviates plus autocorrelation (variogram with range = 15). Following Eq. 3, five more species were created, with random normal deviates, spatial autocorrelation, plus an effect of an environmental variable on each of the five species, with transfer parameter $\beta = 0.5$. The preliminary species were then transformed to species presence-absence or abundance data, as described in the "Data generation" subsection. The random N(0,1.5) constants multiplying the species to make their means different were –3.21, 2.07, –1.06, –3.39, 2.54, 1.68, –1.70, 1.44, 1.48, and 0.11. 50.4% of the entries in the two species files were zeros.

PCNM base functions were created using the program SpaceMaker2 (Borcard and Legendre 2004). We asked the program to create a regular (10 x 10) grid and used 1.4143 as the truncation distance, meaning that the west-east, south-north, and diagonal connections between adjacent points were preserved. In view of canonical partitioning, forward selection of PCNM base functions was carried out in Canoco v. 4.5 (ter Braak and Smilauer 2002). Nine PCNM variables were retained for each data set (species presence-absence and abundance, Hellinger-transformed), as shown in the notes of Table B1. Variation partitioning was done using a canonical partitioning program written by PL; identical results would have been obtained from simple and partial RDA in Canoco. Prior to Mantel-type partitioning, distance matrices were computed as described in the simulation study; Hellinger distances were computed for the two raw species data sets. Partitioning was conducted using the program Permute! version 3.4 (Casgrain 2001).

Supplement 2 contains the two raw species data files (Species_pres-abs.txt and Species_abund.txt), the same files after Hellinger transformation (Species_pres-abs_Hell.txt and Species_abund_Hell.txt) which were used to carry out the redundancy analyses reported in Table

B1, the environmental data file (Envir.txt), the file with the geographic coordinates of the sites (CoordXY.txt), as well as the two files of PCNM base functions after forward selection (9_PCNMs_for_pres-abs.txt and 9_PCNMs_for_abundance.txt), as described in the previous paragraph.

The results for this data set illustrate the first result of the simulation study: compared to canonical partitioning, regression on distance matrices highly underestimates the portions of the species variation explained by the environmental variables alone (Table B1 columns [a+b]), the spatial relationships alone (columns [b+c]), or both sources of variation (columns [a+b+c]).

For the species abundance data, the partitioning results would have been very different without the Hellinger transformation. With the environmental variables and the 3rd-order polynomial of the geographic coordinates as explanatory variables, for example, the total portion of explained variation [a+b+c] would have been 0.1624 (fraction not significant, p = 0.254), while after Hellinger transformation of the abundance data [a+b+c] = 0.2014 (fraction significant, p = 0.002) as shown in Table B1. For presence-absence data, the result without Hellinger transformation ([a+b+c] = 0.2320, p = 0.001) would have been comparable to the one reported in Table B1 ([a+b+c] = 0.2276, p = 0.001) with Hellinger transformation. The Hellinger transformation has very little effect on this particular presence-absence data set because 75% of the sites have from 4 to 6 species present. If all the sites had the exact same number of species present, the Hellinger transformation would have no effect at all on the presence-absence canonical analysis results.

LITERATURE CITED

Borcard, D. and P. Legendre. 2004. SpaceMaker2 – User's guide. Département de sciences biologiques, Université de Montréal. 20 pages. Available from the WWWeb page <http://www.bio.umontreal.ca/legendre/>.

Casgrain, P. 2001. Permute! version 3.4 – User's manual. Département de sciences biologiques, Université de Montréal. Available from the WWWeb page <http://www.bio.umontreal.ca/legendre/>.

ter Braak, C. J. F. and P. Smilauer. 2002. Canoco reference manual and CanoDraw for Windows user's guide: software for canonical community ordination (version 4.5). Microcomputer Power, Ithaca, New York.

TABLE B1. Example data: portion of the species variation in the fractions defined in Fig. 1. The environmental variables were included in all analyses, in the form of either a raw data table (RDA) or a distance matrix (Mantel). "Mantel" refers to regression on distance matrices.

Partitioning method	[a+b+c]	[a+b]	[b+c]	[a]	[b]	[c]	[d]
a) Species presence-absence							
RDA, XY	0.1546[3]	0.1384[3]	0.0648[3]	0.0898[3]	0.0487	0.0162[NS]	0.8455
RDA, polynomial	0.2276[3]	0.1384[3]	0.1543[3]	0.0733[2]	0.0651	0.0892[NS]	0.7724
RDA, PCNM[§]	0.2880[3]	0.1384[3]	0.2043[3]	0.0837[2]	0.0548	0.1495[3]	0.7121
Mantel, D(XY)	0.0184[3]	0.0181[3]	0.0046[1]	0.0139[2]	0.0042	0.0004[NS]	0.9816
Mantel, D(polyn.)	0.0183[3]	0.0181[3]	0.0044[1]	0.0139[2]	0.0041	0.0003[NS]	0.9817
Mantel, ln(D(XY))	0.0191[3]	0.0181[3]	0.0055[2]	0.0135[2]	0.0045	0.0010[NS]	0.9809
b) Species abundance							
RDA, XY	0.1197[2]	0.0982[3]	0.0515[2]	0.0682[NS]	0.0300	0.0215[NS]	0.8803
RDA, polynomial	0.2014[2]	0.0982[3]	0.1414[3]	0.0600[NS]	0.0382	0.1033[NS]	0.7986
RDA, PCNM[§§]	0.2666[3]	0.0982[3]	0.1966[3]	0.0700[1]	0.0282	0.1684[3]	0.7334
Mantel, D(XY)	0.0069[2]	0.0061[2]	0.0031[2]	0.0038[1]	0.0023	0.0009[NS]	0.9931
Mantel, D(polyn.)	0.0070[2]	0.0061[2]	0.0033[2]	0.0037[1]	0.0024	0.0009[NS]	0.9930
Mantel, ln(D(XY))	0.0077[3]	0.0061[2]	0.0040[2]	0.0037[1]	0.0024	0.0016[1]	0.9923

Notes:

[§] 9 PCNM base functions were retained by forward selection (Canoco 4.5) at $p \leq 0.05$: PCNM #1, 3, 4, 6, 9, 16, 41, 42, 57.

[§§] 9 PCNM base functions were retained by forward selection (Canoco 4.5) at $p \leq 0.05$: PCNM #1, 5, 6, 10, 12, 26, 42, 44, 57.

Significance of the fractions (permutation tests, 999 permutations): [1] $0.01 < p \leq 0.05$, [2] $0.001 < p \leq 0.01$, [3] $p \leq 0.001$, [NS] not significant ($p > 0.05$). Fractions [b] and [d] cannot be tested for significance.

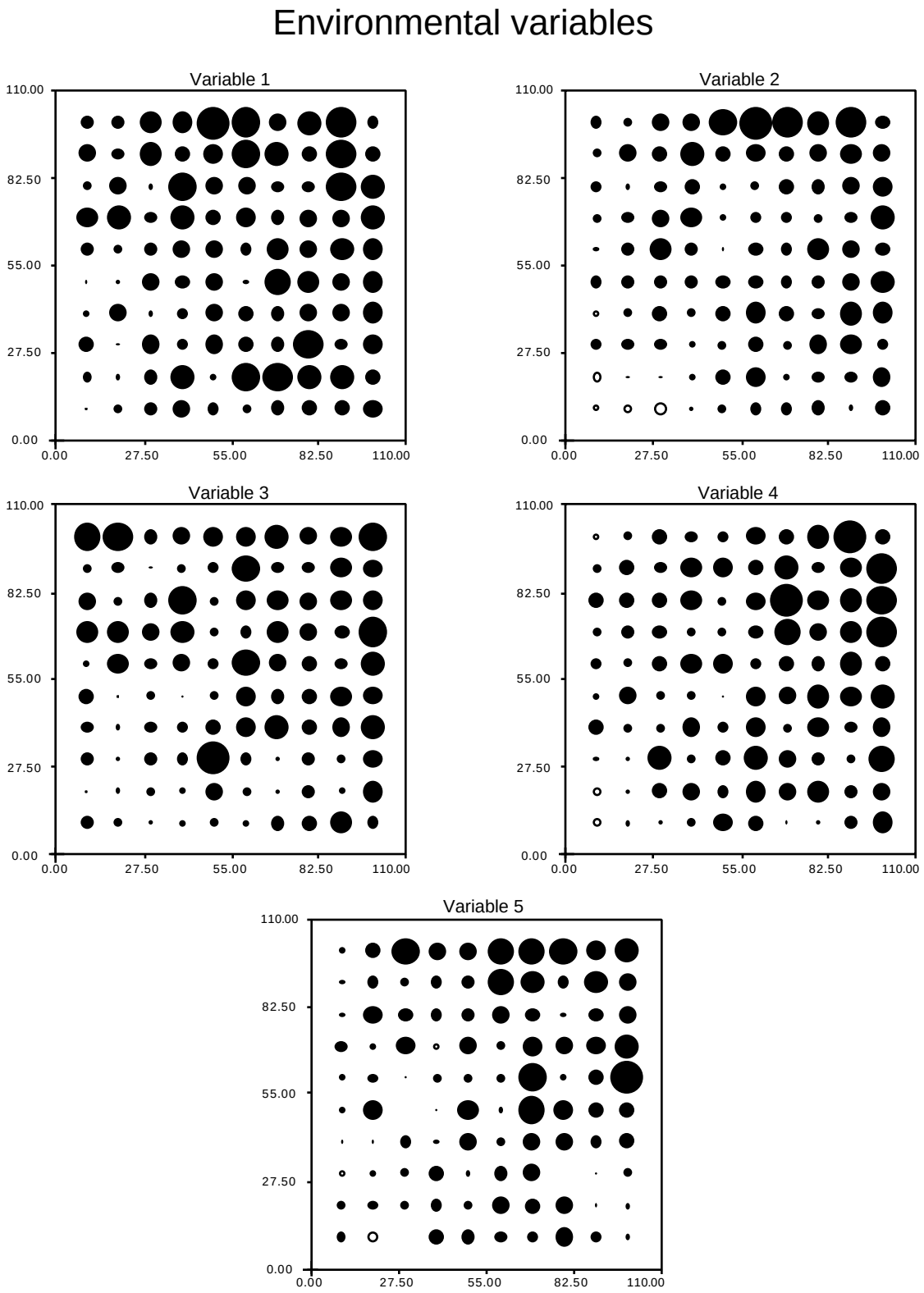

Fig. B1. Values of the five environmental variables at 100 sampling sites on the 10 x 10 regular grid positioned in a 100 x 100 field (arbitrary units). Each variable was generated independently of the others; it contains a deterministic structure (gradient), spatial autocorrelation, and random "innovation" at each site. Dark circles: positive values; empty circles: negative values.

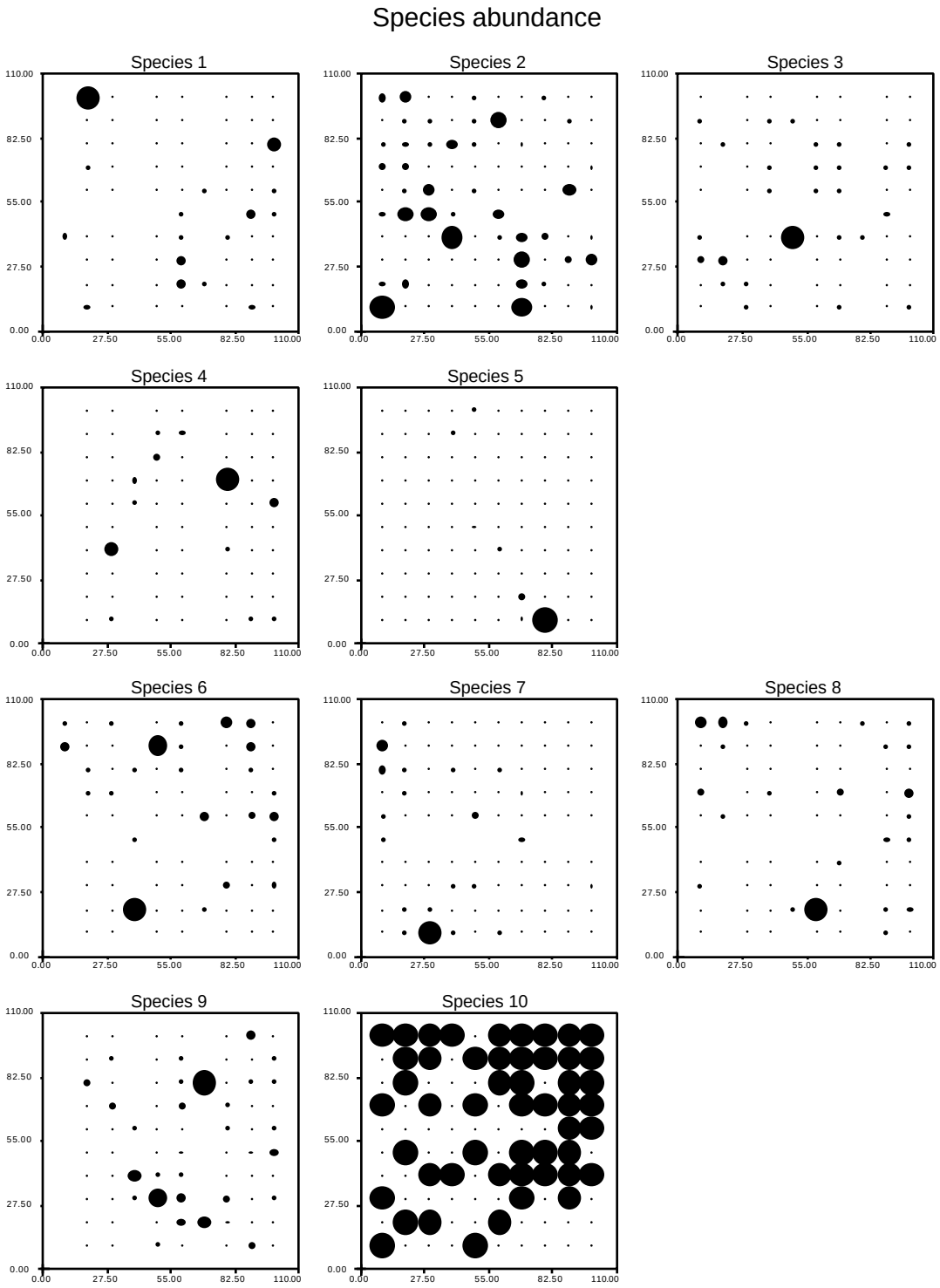

Fig. B2. Species abundance (not Hellinger-transformed) at the 100 sampling sites on the 10 x 10 regular grid positioned in a 100 x 100 field (arbitrary units). Species 1 to 5 are only structured by spatial autocorrelation plus random error; species 6 to 10 are also related to the environmental variables shown in Fig. B1. Raw abundances, before Hellinger transformation. Dark circles: positive values; empty circles: negative values. Bubble sizes are standardized within each graph. They are comparable within a species map but not among maps. Because of its small multiplier, species 10 only contains 0's (49%) and 1's (51%).

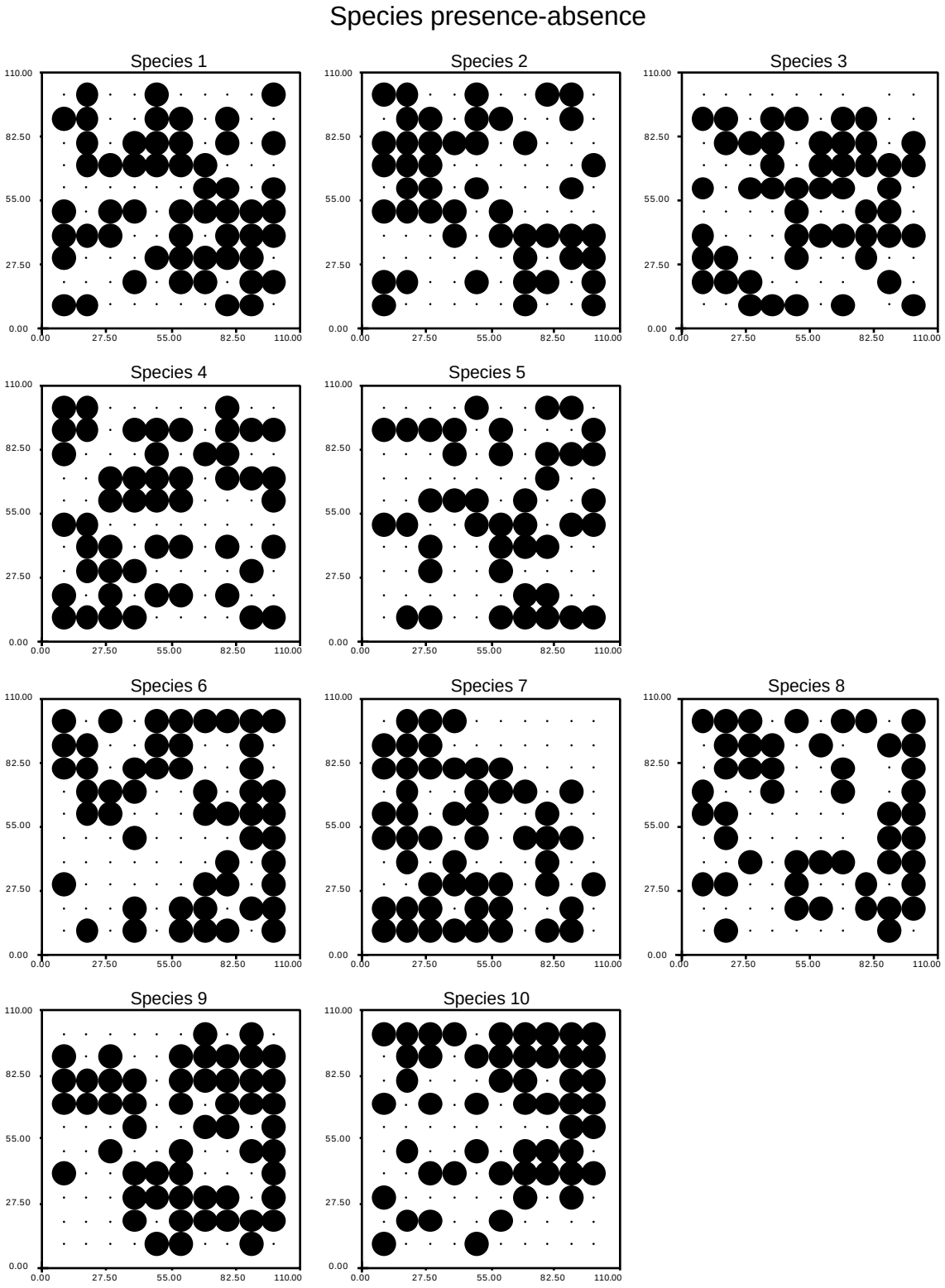

Fig. B3. Species presence-absence data at the 100 sampling sites on the 10 x 10 regular grid positioned in a 100 x 100 field (arbitrary units). Dark circles: species present; points: species absent. Patchiness due to autocorrelation in present in all species. The effect of the gradient present in the environmental variables is visible in the distributions of species 6 to 10, which are each related to one of the environmental variables by a transfer parameter $\beta = 0.5$.

APPENDIX C

Ecological Archives M075-017-A3

SIMULATION RESULTS: PORTIONS OF VARIATION PRODUCED BY THE PARTITIONING METHODS

TABLE C1. Simulation results: portion of the species variation in the fractions shown in Fig. 1 (means after 1000 simulations). Fraction [d], which contains the residuals, is 1 – [a+b+c]. S = species; E = environmental variables; XY = geographic coordinates of the sites. Range = range parameter of the variogram for generation of autocorrelation; the size of the simulation field was 100 x 100 (arbitrary units).

Partitioning method	Species presence-absence						Species abundance					
	[a+b+c]	[a+b]	[b+c]	[a]	[b]	[c]	[a+b+c]	[a+b]	[b+c]	[a]	[b]	[c]
A) S unrelated to E ($\beta = 0$), S autocorrelated (range = 15), E = $N(0,1)$												
RDA, XY	.0756	.0502	.0255	.0501	.0001	.0254	.0757	.0500	.0260	.0497	.0002	.0258
RDA, polynomial	.1586	.0502	.1092	.0494	.0008	.1084	.1603	.0500	.1114	.0489	.0011	.1103
RDA, PCNM	.1630	.0502	.1162	.0468	.0034	.1128	.1743	.0500	.1285	.0458	.0042	.1243
Mantel, D(XY)	.0028	.0015	.0013	.0015	.0000§	.0013	.0018	.0009	.0009	.0009	.0000§	.0009
Mantel, D(polyn.)	.0043	.0015	.0028	.0015	.0000§	.0028	.0024	.0009	.0015	.0009	.0000§	.0015
Mantel, ln(D(XY))	.0027	.0015	.0012	.0015	.0000§	.0012	.0019	.0009	.0010	.0009	.0000§	.0010
B) S related to E ($\beta = 0.5$), S autocorrelated (range = 15), E autocorrelated (range = 15)												
RDA, XY	.1322	.1091	.0252	.1070	.0021	.0231	.1389	.1151	.0259	.1130	.0021	.0238
RDA, polynomial	.2090	.1091	.1085	.1005	.0086	.0999	.2165	.1151	.1107	.1058	.0093	.1014
RDA, PCNM	.2123	.1091	.1174	.0949	.0142	.1032	.2260	.1151	.1277	.0983	.0168	.1109
Mantel, D(XY)	.0059	.0046	.0013	.0046	.0000§	.0013	.0052	.0044	.0008	.0044	.0000§	.0008
Mantel, D(polyn.)	.0075	.0046	.0029	.0046	.0000§	.0029	.0058	.0044	.0014	.0044	.0000§	.0014
Mantel, ln(D(XY))	.0057	.0046	.0011	.0046	.0000§	.0011	.0053	.0044	.0009	.0044	.0000§	.0009
C) S related to E ($\beta = 0.25$), S autocorrelated (range = 15), E autocorrelated (range = 15)												
RDA, XY	.0933	.0689	.0253	.0680	.0009	.0244	.0955	.0703	.0260	.0695	.0009	.0251
RDA, polynomial	.1742	.0689	.1088	.0654	.0035	.1053	.1773	.0703	.1108	.0665	.0038	.1070
RDA, PCNM	.1778	.0689	.1159	.0618	.0071	.1089	.1897	.0703	.1275	.0622	.0081	.1194
Mantel, D(XY)	.0032	.0020	.0012	.0020	.0000§	.0012	.0021	.0013	.0008	.0013	.0000§	.0008
Mantel, D(polyn.)	.0046	.0020	.0026	.0020	.0000§	.0026	.0027	.0013	.0014	.0013	.0000§	.0014
Mantel, ln(D(XY))	.0030	.0020	.0010	.0020	.0000§	.0010	.0022	.0013	.0009	.0013	.0000§	.0009

TABLE C1 (continued)

Partitioning method	Species presence-absence						Species abundance					
	[a+b+c]	[a+b]	[b+c]	[a]	[b]	[c]	[a+b+c]	[a+b]	[b+c]	[a]	[b]	[c]
D) S related to E (β = 0.5), S autocorrelated (range = 15), E with deterministic gradient												
RDA, XY	.1343	.1129	.0578	.0765	.0364	.0214	.1412	.1197	.0609	.0803	.0395	.0214
RDA, polynomial	.2117	.1129	.1378	.0739	.0390	.0988	.2198	.1197	.1426	.0772	.0425	.1001
RDA, PCNM	.2241	.1129	.1487	.0754	.0374	.1113	.2360	.1197	.1581	.0781	.0416	.1163
Mantel, D(XY)	.0110	.0090	.0054	.0056	.0034	.0020	.0112	.0096	.0055	.0057	.0039	.0016
Mantel, D(polyn.)	.0115	.0090	.0037	.0078	.0012	.0025	.0109	.0096	.0029	.0080	.0015	.0014
Mantel, ln(D(XY))	.0111	.0090	.0054	.0057	.0034	.0020	.0113	.0096	.0055	.0058	.0038	.0017
E) S related to E (β = 0.25), S autocorrelated (range = 15), E with deterministic gradient												
RDA, XY	.0928	.0705	.0347	.0581	.0124	.0223	.0952	.0729	.0361	.0591	.0137	.0224
RDA, polynomial	.1750	.0705	.1182	.0568	.0137	.1045	.1787	.0729	.1210	.0577	.0152	.1058
RDA, PCNM	.1860	.0705	.1286	.0574	.0131	.1155	.1951	.0729	.1373	.0578	.0151	.1222
Mantel, D(XY)	.0040	.0025	.0019	.0021	.0004	.0015	.0032	.0021	.0017	.0015	.0006	.0011
Mantel, D(polyn.)	.0052	.0025	.0028	.0024	.0002	.0026	.0036	.0021	.0018	.0018	.0003	.0015
Mantel, ln(D(XY))	.0039	.0025	.0019	.0020	.0005	.0014	.0034	.0021	.0019	.0015	.0007	.0012

Note:
[§] In Mantel partitioning results, [b] fractions often had their first non-zero digits at the 5th or 6th decimal place.

SUPPLEMENT 1

Ecological Archives M075-017-S1

DATA GENERATION AND ANALYSIS PROGRAMS USED IN THE SIMULATION STUDY

This Supplement contains the source code and executables for SIMSSD4, a FORTRAN program used for generation of the species, environment, and geographic coordinate data files in the simulation study, as well as the MATLAB code used to automatically carry out canonical and Mantel variation partitioning, with permutation tests, on the generated data sets. The following files are available in ESA's *Ecological Archives* M075-017-S1:

SimSSD program

• SimSSD4_g77.for: Fortran source code for SimSSD for Macintosh OS X and Windows.

• SimSSD4 user's guide.pdf: SimSSD user's manual.

• File_of_parameters.txt: file of parameters for test run.

• Coord.txt, Species.txt, Envir.txt: results of the test run of SimSSD.

All these files, as well as executables for Macintosh OS X and Windows, are available in the compressed files SimSSD_for_OS_X.zip and SimSSD_for_Windows.zip.

MATLAB functions

• Guide.pdf: guide to the 8 MATLAB functions and 5 text files containing the necessary files to run one of the scenarios considered in our study.

• MatlabFunctions.zip: contains all the files described in the Guide.pdf document.

SUPPLEMENT 2

Ecological Archives M075-017-S2

DATA TABLES USED IN THE EXAMPLE DETAILED IN APPENDIX B

This Supplement contains the data tables used in the example detailed in Appendix B. The raw species and environmental data were generated by the program SimSSD.

File list

The compressed folder Supplement_2.zip, available in ESA's *Ecological Archives* M075-017-S2, contains the following files:

• Species_pres-abs.txt: raw presence-absence species data.

• Species_pres-abs_Hell.txt: same after Hellinger transformation.

• Species_abund.txt: raw abundance species data.

• Species_abund_Hell.txt: same after Hellinger transformation.

• Envir.txt: environmental data.

• CoordXY.txt: geographic coordinates of the sites.

• 9_PCNMs_for_pres-abs.txt: 9 PCNM base functions selected by forward selection against the Hellinger-transformed presence-absence species data.

• 9_PCNMs_for_abundance.txt: 9 PCNM base functions selected by forward selection against the Hellinger-transformed abundance species data.

Data generation using SimSSD

The raw species and environmental data were generated by the program SimSSD4 using the following line of of parameters:

1 100 100 2 5 0.5 0 15 15 0 15 15 1 2 100 5 5

The answers to the questions of the program were the following:

Transformation of species data: 3 for presence-absence, 1 for abundance data

Standard deviation of the random normal species multipliers: 1.5

Type a small positive integer to offset the random number generator for generation of data: 1