

Beta diversity as the variance of community data: dissimilarity coefficients and partitioning

Pierre Legendre^{1*}, Miquel De Cáceres^{2,3}

¹ *Département de sciences biologiques, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Québec, Canada H3C 3J7*

² *Centre Tecnològic Forestal de Catalunya. Ctra. St. Llorenç de Morunys km 2, 25280 Solsona, Catalonia, Spain*

³ *CREAF (Centre de Recerca Ecològica i Aplicacions Forestals), Bellaterra, Catalonia, Spain*

E-mail addresses: pierre.legendre@umontreal.ca, miquelcaceres@gmail.com

Short running title: Beta diversity partitioning

Keywords: beta diversity, community ecology, community composition data, dissimilarity coefficients, local contributions to beta diversity, properties of dissimilarity coefficients, species contributions to beta diversity, variance partitioning

Article type: Ideas and Perspectives

Words in abstract: 200

Words in main text: 7500

Text boxes: 0

References: 61

Figures: 4

Tables: 2

*Correspondence

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

Phone: 514-343-7591
Fax: 514-343-2293
E-mail: Pierre.Legendre@umontreal.ca

Authorship statement

The two authors contributed equally to the paper and took the lead at different times. PL coordinated the writing and editing of the final version of the manuscript.

Abstract

Beta diversity can be measured in different ways. Among these, the total variance of the community data table \mathbf{Y} can be used as an estimate of beta diversity. We show how the total variance of \mathbf{Y} can be calculated either directly or through a dissimilarity matrix obtained using any dissimilarity index deemed appropriate for pairwise comparisons of community composition data. We addressed the question of which index to use by coding 16 indices using 14 properties that are necessary for beta assessment, comparability among data sets, sampling issues, and ordination. Our comparison analysis classified the coefficients under study into five types, three of which are appropriate for beta diversity assessment. Our approach links the concept of beta diversity with the analysis of community data by commonly used methods like ordination and ANOVA. Total beta can be partitioned into Species Contributions (SCBD: degree of variation of individual species across the study area) and Local Contributions (LCBD: comparative indicators of the ecological uniqueness of the sites) to Beta Diversity. Moreover, total beta can be broken up into within- and among-group components by MANOVA, into orthogonal axes by ordination, into spatial scales by eigenfunction analysis, or among explanatory data sets by variation partitioning.

INTRODUCTION

A most interesting property of species diversity is its organization through space. This phenomenon, now well known to community ecologists, was first discussed by Whittaker in two seminal papers (1960, 1972) where he described the alpha, beta and gamma diversity levels of natural communities. Alpha is local diversity, beta is spatial differentiation, and gamma is regional diversity. The interest of community ecologists for beta diversity stems from the fact that spatial variation in species composition allows them to test hypotheses about the processes that generate and maintain biodiversity in ecosystems. Sampling through space, time, or along gradients representing processes of interest is a way of carrying out *mensurative experiments* (Hurlbert 1984) involving natural processes without the constraints (e.g. small sample size) of controlled experiments.

Beta diversity is conceptually the variation in species composition among sites within a geographic area of interest (Whittaker 1960). Several authors have used that description of the concept, including Legendre *et al.* (2005), Anderson *et al.* (2011) and Baselga & Orme (2012). Different equations have been proposed to measure that variation. Vellend (2001) and Anderson *et al.* (2011) pointed out that studies of beta diversity might focus on two aspects of community structure, distinguishing two types of beta diversity. The first is turnover, or the directional change in community composition from one sampling unit to another along a predefined spatial, temporal, or environmental gradient. The second is variation in community composition among sampling units, which is a non-directional approach because it does not make reference to any explicit gradient. Both approaches are legitimate.

Regardless of whether beta diversity is defined as directional or non-directional, one can be interested in summarizing it using a single number that quantifies the variation. A lot of interest has been centred on the choice of the best index to produce that number. In the directional approach, the slope of the similarity decay in species composition with geographical distance can be used as a measure of beta (Nekola & White 1999). In his 1960 paper, Whittaker suggested to compute a non-directional beta index for species richness as $\beta = \gamma/\alpha$ where γ is the number of species in the region and α is the mean number of species at the study sites within the region. Since then, several other indices have been suggested to estimate a value corresponding to beta in the turnover and non-directional frameworks; see Vellend (2001), Koleff *et al.* (2003) and Anderson *et al.* (2011) for reviews. Currently, the most popular indices belong to two families that can be labelled the additive ($H_\alpha + H_\beta = H_\gamma$) and multiplicative ($H_\alpha \times H_\beta = H_\gamma$) approaches (Jost 2007, Chao *et al.* 2012). A detailed discussion of these two families is found in a *Forum* section published by *Ecology* (2010:1962–1992).

In his introduction to the *Forum*, Ellison (2010) noted that in the additive and multiplicative approaches, beta is a derived quantity that is numerically related to alpha and gamma. He pointed out that it would be most useful to have a method to estimate beta diversity without prior computation of alpha and gamma; he called for computational independence, which does not imply statistical independence. The approach adopted and developed in the present paper is to use the total variance of the site-by-species community table \mathbf{Y} as a single-number estimate of beta diversity (Pelissier *et al.* 2003, Legendre *et al.* 2005, Anderson *et al.* 2006). Fulfilling Ellison's wish, it is computed without reference to the values of alpha and gamma and its statistical dependence on gamma can be accounted for using null models (Kraft *et al.* 2011, De Cáceres *et al.* 2012). While acknowledging that other measures of beta can also achieve computational and

statistical independence (e.g., Chao *et al.* 2012), one of our aims is to stress an important advantage of the total variance of \mathbf{Y} over other measures: it allows ecologists to go beyond the single-number approach and partition the spatial variation in several ways to answer precise ecological questions and test hypotheses about the origin and maintenance of beta diversity in ecosystems.

We will explore the advantages and limitations of estimating beta diversity (BD_{Total}) as the total variation of the community matrix \mathbf{Y} . (1) In a first section, we show that BD_{Total} can be obtained in two equivalent ways, i.e. by computing the sum of squares of the species occurrence or abundance data or *via* a dissimilarity matrix. When the first method is used, species abundances should be transformed in an appropriate way before computing BD_{Total} . The second method is also appealing because it allows the estimation of beta using the dissimilarity functions that are appropriate for the analysis of community data. (2) There are, however, many different dissimilarity coefficients, and not all of them are appropriate for estimating beta diversity. A comparative analysis of 16 coefficients is undertaken in the next section to guide users faced with the problem of choosing a coefficient. (3) We then present an example to illustrate the calculation of beta as the total variance of \mathbf{Y} and the contributions of individual species and sampling units. (4) Following that, we show that the proposals of Whittaker (1972) and Ricotta & Marignani (2007) are special cases of BD_{Total} computed from a dissimilarity matrix, and that the beta diversity statistic of Anderson *et al.* (2006) is closely related to BD_{Total} . (5) Finally, we show that the total variance of \mathbf{Y} links beta diversity assessment with the description (through ordination) and hypothesis testing (through regression and canonical analysis) phases of community ecology, as well as other variance partitioning methods.

BETA DIVERSITY AS THE TOTAL COMMUNITY COMPOSITION VARIANCE

Equivalent ways of computing $\text{Var}(\mathbf{Y})$

This section presents two equivalent ways of computing the total variance of the community composition matrix \mathbf{Y} . The first one is straightforward, it is simply the total variance of matrix \mathbf{Y} . The second one is based upon community dissimilarity matrices computed using the indices developed by ecologists over more than a century. The section also shows that the total variance can be divided into the contributions of individual species and individual sampling sites. Readers can follow the explanation on the diagram in Fig. 1.

Let $\mathbf{Y} = [y_{ij}]$ be a data table containing the presence-absence or the abundance values of p species (column vectors $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_p$ of \mathbf{Y}) observed in n sampling units (row vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ of \mathbf{Y}). We will use indices i and h for sampling units, index j for species, and y_{ij} for individual values in \mathbf{Y} . The total variance of \mathbf{Y} , noted $\text{Var}(\mathbf{Y})$, can be computed as follows:

I. Sums of squares. — The usual way to obtain $\text{Var}(\mathbf{Y})$ consists in computing a matrix of squared deviations from the column means. Let \mathbf{S} (for “square”) be a $n \times p$ rectangular matrix where each element s_{ij} is the square of the difference between the y_{ij} value and the mean value of the corresponding j th species:

$$s_{ij} = (y_{ij} - \bar{y}_j)^2. \quad (1)$$

All s_{ij} values in column j are zero if all sites have the same abundance for species j . If we sum all values of \mathbf{S} , we obtain the total sum of squares (SS) of the species composition data:

$$\text{SS}_{\text{Total}} = \sum_{i=1}^n \sum_{j=1}^p s_{ij}. \quad (2)$$

This quantity forms the basis of BD_{Total} , which is the index of beta diversity whose properties are studied in this paper:

$$BD_{\text{Total}} = \text{Var}(\mathbf{Y}) = SS_{\text{Total}} / (n - 1) . \quad (3)$$

Equation 3 converts the sum of squares into the usual unbiased estimator of the variance, whose values can be compared between data matrices having different numbers of sampling units. SS_{Total} and $\text{Var}(\mathbf{Y}) = BD_{\text{Total}}$ were both proposed by Legendre *et al.* (2005) as measures of beta diversity. The two indices are equally useful to compare repeated surveys of a region involving the same sites, or for simulation studies, but there is a clear advantage in using $\text{Var}(\mathbf{Y})$ for comparisons among regions.

Although we advocate using $\text{Var}(\mathbf{Y})$ as a measure of beta diversity, it is important to note that eqs 1-3 should not be computed directly on raw species abundance or biomass data. Because calculating $\text{Var}(\mathbf{Y})$ on raw species abundances entails that the dissimilarity between sites is assessed using the Euclidean distance (eq. 7) and this coefficient is not appropriate for compositional data (see section “Dissimilarity coefficients and beta assessment”), species abundance data should be transformed in an ecologically meaningful way before BD_{Total} is calculated using eqs 1-3.

An advantage of conceiving beta as the total variation in \mathbf{Y} is that SS_{Total} allows the assessment of the *contributions of individual species* and of *individual sampling units to the overall beta diversity*. That is, one can compute the sum of squares corresponding to the j th species,

$$SS_j = \sum_{i=1}^n s_{ij}^2 \quad (4a)$$

which is the contribution of species j to the overall beta diversity. SS_j divided by $(n - 1)$ is the variance of species j . The *relative* contribution of species j to beta, which we call *Species Contribution to Beta Diversity* (SCBD), is thus:

$$SCBD_j = SS_j / SS_{\text{Total}} \quad (4b)$$

In an analogous way, one can compute the sum of squares corresponding to the i th sampling unit,

$$SS_i = \sum_{j=1}^p s_{ij}^2 \quad (5a)$$

The SS_i values represent a genuine partitioning of beta diversity among the sites. Because the s_{ij} values are squared deviations from the species means, SS_i is the squared distance of sampling unit i to the centroid of the distribution of sites in species space. SS_i also measures the leverage of site i in a PCA ordination. The *relative* contribution of sampling unit i to beta diversity, which we call *Local Contribution to Beta Diversity* (LCBD _{i}), is thus:

$$LCBD_i = SS_i / SS_{\text{Total}} \quad (5b)$$

LCBD values can be mapped, as will be shown in the ecological illustration below. Ecologically, they represent *the degree of uniqueness of the sampling units in terms of community composition*. Mapping the centred values using different symbols or colours is a way to highlight the sites with LCBD values higher and lower than the mean.

LCBD indices can be tested for significance by random, independent permutations within the columns of matrix \mathbf{Y} ; testing the LCBD _{i} is the same as testing the SS_i indices. This permutation method tests H_0 that the species are distributed at random, independently of one another, among the sites, while preserving the species abundance distributions found in the observed data. However, it destroys the association of the species to the site ecological

conditions, as well as the spatial structure of community composition resulting from assembly processes (e.g. dispersal, environmental filtering, etc.). Note that the species richness (alpha diversity) of the sites is changed by this permutation method; species-poor sites become richer in most permutations and species-rich sites become poorer. Arguably, these two kinds of sites may have large LCBD for that reason, so this permutation method includes randomization of species richness in its null hypothesis. Other null hypotheses may be tested using other permutation schemes, for example by preserving site attributes such as total species richness or number of individuals (e.g. in De Cáceres *et al.* 2012). A simulation study that we performed showed that the LCBD test described here has correct rates of type I error for all coefficients that are suitable for beta diversity study (identified in section “Comparative study”).

Hence, the two decompositions of SS_{Total} are:

$$SS_{\text{Total}} = \sum_{j=1}^p SS_j \quad \text{and} \quad SS_{\text{Total}} = \sum_{i=1}^n SS_i . \quad (6a,b)$$

2. Dissimilarity. — As mentioned above, there is an alternative path starting from \mathbf{Y} and leading to SS_{Total} (Fig. 1). That is, SS_{Total} can also be obtained from an $n \times n$ symmetric dissimilarity matrix $\mathbf{D} = [D_{hi}]$ containing Euclidean distances among points, computed using the classical Euclidean distance formula:

$$D_{hi} = D(\mathbf{x}_h, \mathbf{x}_i) = \sqrt{\sum_{j=1}^p (y_{hj} - y_{ij})^2} . \quad (7)$$

The following equivalence is described in Legendre *et al.* (2005) and in Legendre & Legendre (2012, Chapter 8):

$$SS_{\text{Total}} = \frac{1}{n} \sum_{h=1}^{n-1} \sum_{i=h+1}^n D_{hi}^2 . \quad (8)$$

That is, one can obtain SS_{Total} by summing the squared distances in the upper or lower half of matrix \mathbf{D} and dividing by the number of objects n (*not* by the number of distances). This equality (eq. 8) is demonstrated in Appendix 1 of Legendre & Fortin (2010).

The Euclidean distance has long been known to be inappropriate for the analysis of community composition data (see next section). For that reason, equations 7-8 should not be used to compute SS_{Total} unless species abundance data have been appropriately transformed so that the resulting dissimilarity assessments are ecologically meaningful (e.g. using the Hellinger or chord transformations described in Appendix S1 in *Supporting Information*). Equation 8 can also be generalized to distance matrices obtained using other dissimilarity indices. These indices may or may not have the Euclidean property (P13 below), but their other properties may make them appropriate for beta diversity assessment. Thus, a valid method to calculate BD_{Total} consists in computing a dissimilarity matrix \mathbf{D} using a selected ecological dissimilarity coefficient instead of the Euclidean distance, and applying eq. 8 to obtain SS_{Total} , followed by eq. 3. That eq. 8 applies to ecological dissimilarities that have the Euclidean property, or not, is shown in Appendix S2. How to choose an appropriate dissimilarity coefficient for a given study is described in the next section.

It is possible to calculate the contributions of individual sampling units from \mathbf{D} . Indeed, the algebra of principal coordinate analysis (PCoA, Gower 1966) offers a way of computing the sum of squares SS_i , corresponding to each sampling unit i , directly from \mathbf{D} . In PCoA, prior to eigen-decomposition, the distance matrix is transformed into matrix $\mathbf{A} = [a_{hi}] = [-0.5D_{hi}^2]$, then centred as proposed by Gower (1966) using the equation

$$\mathbf{G} = \left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n} \right) \mathbf{A} \left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n} \right) \quad (9)$$

where \mathbf{I} is an identity matrix of size n , $\mathbf{1}$ is a vector of ones (length n) and $\mathbf{1}'$ is its transpose (Legendre & Legendre 2012, eqs 9.40 and 9.42). The diagonal elements of matrix \mathbf{G} are the SS_i values, or the squared distances of the points to the multivariate centroid of \mathbf{Y} , which is located at the centroid of the principal coordinate space:

$$[SS_i] = \text{diag}(\mathbf{G}) . \quad (10a)$$

The vector of local contributions of the sites to beta diversity ($LCBD_i$) is computed as follows:

$$[LCBD_i] = \text{diag}(\mathbf{G})/SS_{\text{Total}} . \quad (10b)$$

Despite its advantages, working from matrix \mathbf{D} instead of the matrix of squared centred values \mathbf{S} entails the drawback that one loses track of the species. Because \mathbf{D} is computed among sampling units over all species, the contributions of individual species cannot be recovered from \mathbf{D} .

To summarize:

- One can compute the total sum of squares in the community data \mathbf{Y} , SS_{Total} , from either the community composition matrix \mathbf{Y} (eq. 2) or from a Euclidean distance matrix \mathbf{D} among sites (eq. 8). The two modes of calculation produce the same statistic, SS_{Total} , and from it one can compute the total variance, $BD_{\text{Total}} = \text{Var}(\mathbf{Y})$ (eq. 3). In both cases, the community data table \mathbf{Y} should have been transformed in an appropriate way for the results to be meaningful in terms of beta diversity assessment.
- Alternatively, one can use eq. 8 to compute SS_{Total} from a dissimilarity matrix \mathbf{D} obtained using any appropriate dissimilarity coefficient (next section). Equation 8 applies to ecological dissimilarity indices that have the Euclidean property, or not, as demonstrated in Appendix S2.

- The contribution of the i th sampling unit to the overall beta diversity can be computed using eq. 5a. From these, *Local Contribution to Beta Diversity* (LCBD) coefficients can be derived. LCDB coefficients are comparative indicators of the ecological uniqueness of the sites in terms of community composition. The SS_i values are also found on the diagonal of matrix **G** (eqs 9 and 10a). The relative contributions (LCDB) are computed using eqs 5b and 10b.
- If BD_{Total} is calculated from **Y** (eq. 3) transformed in an appropriate way, the contribution of species j to the overall beta diversity, SS_j , is computed using eq. 4a, and the relative contributions, called the *Species Contributions to Beta Diversity* (SCBD), are computed using eq. 4b. SCBD coefficients represent the degree of variation of individual species across the study area. SS_j and SCDB coefficients cannot be derived from a distance matrix.

DISSIMILARITY COEFFICIENTS AND BETA ASSESSMENT

Since the description of the first floristic similarity coefficient by Paul Jaccard (1900), community ecologists have developed a broad array of similarity and dissimilarity coefficients. Ecologists are often faced with the question: Which community data transformation and/or (dis)similarity coefficient should I use in my study? When assessing beta diversity through the variation in community composition, one needs to specify what is meant by “variation in community composition”. The answer will determine the choice of a community data transformation and/or dissimilarity measure, and must be carefully articulated (Anderson *et al.* 2006).

There is no single coefficient that is appropriate in all occasions. Choice should be guided by the properties of coefficients and the objective of the research. Several studies have compared resemblance coefficients, focussing on their linearity and resolution along simulated gradients

(e.g. Bloom 1981, Hajdu 1981, Gower & Legendre 1986, Faith *et al.* 1987, Legendre & Gallagher 2001), or investigating theoretical properties (e.g. Janson & Vegelius 1981, Hubálek 1982, Wilson & Shmida 1984, Gower & Legendre 1986, Koleff *et al.* 2003, Chao *et al.* 2006, Clarke *et al.* 2006). Complementing these studies, we present in this section a comparative review of several abundance- and incidence-based dissimilarity coefficients, listed in Table 1. Our aim is to determine which coefficients are the most appropriate for assessing beta diversity under the present approach. We restricted the list to the coefficients originally designed for pairwise comparisons, thus excluding multiple-site measures (e.g. Baselga 2010). In addition, we focused on properties that are easy to understand and interpret ecologically, with preference for those that could be checked unequivocally.

Properties of dissimilarity indices for the study of beta diversity

Fourteen properties, divided in four groups, are described in Appendix S3, which also outlines procedures to check which dissimilarity indices possess them. The first two groups (P1 to P9) contain the minimum requirements for assessing beta diversity. The remaining two groups (P10 to P14) are not necessarily required in all beta diversity studies. Practitioners should determine whether the context of their analyses requires these properties, or not. Other properties are also considered interesting by authors of other studies on dissimilarity coefficients.

The dissimilarity coefficients

A selection of 16 quantitative dissimilarity coefficients commonly used for beta diversity assessment was considered in our comparison study. They represent a broad hand among the available coefficients. Equations are shown in Table 1 for community composition abundance and for presence-absence (i.e. incidence) data. Table 2 indicates which dissimilarity coefficients

possess the properties mentioned in the previous paragraph and described in Appendix S3, as well as their maximum values (D_{\max}) when they exist.

The first coefficient in the list is the Euclidean distance. Although this distance is known to be inappropriate for the analysis of community composition data sampled under varying environmental conditions (Orlóci 1978, Legendre & Gallagher 2001), it is included in the comparison where it will serve as a reference point. It is the failure of the Euclidean distance to correctly account for beta diversity (it lacks properties P4, P5, P7, P8, P9) that makes it necessary for ecologists to rely on the other dissimilarity measures investigated in this paper. The Euclidean distance may, however, become appropriate after transformation of the community data (Appendix S1). Likewise, the Manhattan distance is inappropriate *per se*; nevertheless, it is included in the comparison because it becomes the Whittaker dissimilarity after profile transformation of \mathbf{Y} , and that index is appropriate for beta diversity studies (Whittaker 1952, Faith *et al.* 1987; Appendix S1).

The other coefficients included in the comparative study are double-zero asymmetric (property P4); they have been recommended and used for community composition assessment or beta diversity studies. Four of these dissimilarities can be computed using the formula in Table 1 or through the alternative method corresponding to property P14. For the species profile, Hellinger, chord, and chi-square distances, the data are first transformed using the same-name transformation (Appendix S1); computing the Euclidean distance (eq. 7) on the transformed data produces the targeted profile, Hellinger, chord, or chi-square distance.

When applied to presence-absence data, several quantitative coefficients in Table 1 produce either the one-complement of the Jaccard similarity index or the one-complement of the Sørensen index. The Hellinger and chord distances both produce $D = \sqrt{2(1 - \text{Ochiai similarity})}$.

Comparative study

The properties of the selected coefficients were coded into a data matrix with the coefficients as rows and properties P4 to P14 as columns (Table 2). Most properties were coded as presence-absence (0-1), except for P13 which was coded on a semiquantitative 0-1-2 scale (0 = not Euclidean, 1 = $\mathbf{D}^{(0.5)}$ is Euclidean, 2 = $\mathbf{D}^{(0.5)}$ and \mathbf{D} are Euclidean). The missing value in Table 2 (coded “NA”) was transformed to 1; the reason is that the chi-square distance has property P7, so it would likely have P10 if a binary form was available for that coefficient. The data matrix was subjected to principal component analysis (PCA) of the correlation matrix.

The analysis produced an ordination of the dissimilarities (Fig. 2) where similar coefficients are close to one another and dissimilar ones are more distant. Properties P4 to P14, which are the variables of the matrix subjected to PCA, are shown as red arrows. One can identify five types of coefficients using the data in Table 2 and the ordination diagram:

- Type I contains the Euclidean and Manhattan distances, as well as the mean character difference and the species profile distance. They all lack several of the important properties in the first two classes (P4 to P9). Most notably, the Euclidean and Manhattan distances do not have the double-zero asymmetry property (P4), and the four coefficients fail to give the largest dissimilarity values to pairs of sites without species in common (P5). The distance between species profiles decreases when the number of unique species in the compared sites increases (P6). The Euclidean distance, Manhattan distance and species profile distances are not species-replication invariant (P7). Moreover, the Euclidean, Manhattan, and modified mean character difference do not fulfil P8 and P9. The species profile distance is invariant to the measurement units of the data (P8), but the upper bound of $\sqrt{2}$ is only reached when there is a single, unique species per site; with more

species, the maximum distance decreases with the number of unique species. Due to these shortcomings, the four coefficients belonging to type I do not allow proper assessment and comparison of beta diversity estimates among data sets.

Coefficients in types II to IV provide asymmetrical treatment of double-zeros (P4) and they all have properties P5 to P9, which are required for comparability of beta estimates among data sets. They are thus all appropriate for beta diversity assessment.

- Type II contains the Hellinger and chord distances. These two distances are closely related: the Hellinger distance is equal to the chord distance computed on square-root-transformed species frequencies. They share all properties in classes 1 and 2, which are necessary for beta diversity assessment. Furthermore, type II coefficients are Euclidean (P13) and they can be emulated by transformations of the raw frequency or biomass data (P14). Hence, **D** matrices computed using these coefficients are fully suitable for ordination by principal coordinate analysis (PCoA), which will not produce negative eigenvalues and complex axes. For an easier and more informative ordination, species frequency (or frequency-like, such as biomass) data transformed using the Hellinger and chord transformations (Appendix S1) can be analysed directly by principal component analysis (PCA) and by canonical redundancy analysis (RDA); this is not the case for the type III and IV coefficients. (PCoA of Hellinger and chord distance matrices produces the same ordinations as PCA of the Hellinger and chord transformed data.) Moreover, SS_{Total} corresponding to the Hellinger and chord distances can be obtained by computing the transformation in Appendix S1, then applying eqs 1 and 2 to the transformed data. This is simpler than computing the distance matrix and using eq. 8 to obtain SS_{Total} . Furthermore, the Hellinger and chord transformed data allow the computation of SCBD statistics (eq. 4b), which cannot be obtained from a distance matrix.

• Type III contains the divergence, Canberra, Whittaker, percentage difference (*alias* Bray-Curtis), Wishart and Kulczynski dissimilarities. They share all properties in classes 1 and 2, which are necessary for beta diversity assessment. The coefficient of divergence, which is Euclidean, can be used directly in PCoA ordination. For four coefficients (Canberra, Whittaker, percentage difference and Wishart), the square root of the distances must be taken before they are used in PCoA. The matrix of principal coordinates can be used as the response data in RDA; this is the distance-based RDA method proposed by Legendre & Anderson (1999). Among the six coefficients in this group, only the Whittaker index is invariant to the total abundance of each sampling unit (P11); the remaining indices are thus affected to some extent by differences in total abundances between the two compared sites. The Kulczynski coefficient is suitable for beta diversity assessment, but not for ordination, and it does not correct for undersampling.

Considering the properties analysed in this paper, this coefficient does not offer any particular advantage not available in other coefficients, and its application is limited.

• Type IV contains the abundance-based quantitative forms of the Jaccard, Sørensen and Ochiai indices. Like coefficients of type II, type IV coefficients fulfil property P11 (invariance to total abundance in individual sampling unit). In addition, they have property P12 (correction for undersampling), but not properties P13 and P14, which are desirable for ordination. In particular, type IV coefficients are not Euclidean (P13) in quantitative form, although the Jaccard, Sørensen and Ochiai similarities, which are the binary counterparts of the first three, produce coefficients with the Euclidean property when transformed to $D = \sqrt{1 - \text{similarity}}$ (Legendre & Legendre 2012, Table 7.2).

• The chi-square distance forms type V. This distance is widely used to analyse communities since it is the basis for correspondence analysis. The chi-square distance gives more importance

to rare than common species in the assessment of the distance between sites, the rare species (when their abundances are correctly estimated by sampling) being considered as more important indicators of special environmental conditions prevailing at some sites. Unfortunately, it lacks property P5, and this makes it unsuitable for beta diversity studies.

Maximum value of BD

All dissimilarities in types II to IV have a maximum value, reached when two sites have completely different community compositions. For example, the Hellinger and chord distances in type II have a minimum value of 0 and a maximum of $\sqrt{2}$ (Table 2). If all sites have entirely different species compositions, all $n(n-1)/2$ distances in **D** are $\sqrt{2}$ and eqs 8 and 3 produce $BD_{Total} = 1$. Hence for these two dissimilarity indices, BD_{Total} is in the range [0, 1]. All other indices that are appropriate for beta assessment (types III and IV) have maximum values of 1. When all sites have different species compositions, the distances are all equal to 1 and BD_{Total} computed through eqs 8 and 3 is 0.5, so that BD_{Total} is in the range [0, 0.5]. For these distances, multiplying BD_{Total} by 2 would directly produce relative BD values (BD_{rel} , Appendix S3, property P9) in the range [0, 1]. Hence BD_{Total} has a fixed range of values for any community, which does not depend on the total abundance in the community composition table.

ECOLOGICAL ILLUSTRATION: FISH BETA DIVERSITY IN DOUBS RIVER

Freshwater fish were collected by Verneaux (1973) in the Doubs River, a tributary of the Saône that runs near the France-Switzerland border in the Jura Mountains in eastern France. In his paper, Verneaux proposed to use fish communities to characterize ecological zones along European rivers and streams. The data include fish community composition at 30 sites along the 453 km course of the river, the site geographic coordinates, and environmental data (source:

<http://www.bio.umontreal.ca/numecolR/>). 27 species were captured and identified. No fish were caught at site 8, hence that site was excluded from the reanalyses made by Borcard *et al.* (2011), as well as here. As in that book, we subjected the fish data to a chord transformation before analysis (Appendix S1).

SS_{Total} (eq. 2) was 15.243 and BD_{Total} (eq. 3) was 0.544 for the fish data. The local contributions of individual sites were computed; the values of SS_i (eq. 5a) ranged from 0.291 to 0.971. An ordination diagram (Fig. 3) illustrates the mathematical meaning of SS_i indices: they are the squares of the distances of the sites to the multivariate centroid, as discussed under eq. 9.

The relative contributions ($LCBD_i = SS_i / SS_{\text{Total}}$, eq. 5b) were in the range [0.019, 0.064]. LCBD indices indicate the uniqueness of the fish community at each site. They are plotted on a schematic map of the river in Fig. 4a, which also shows the two sites where LCBD was statistically significant. Comparison with species richness (Fig. 4b) showed that for this data, LCBD was negatively correlated to richness ($r = -0.60$), indicating that high LCBD (i.e. high uniqueness of species composition) was often related to a small number of species. This is not, however, a general or obligatory relationship.

Environmental variables were also available for each site: distance from the source, altitude, riverbed slope, mean minimum discharge, pH, concentrations in calcium, phosphate, nitrate, ammonium and dissolved oxygen, and biochemical oxygen demand (BOD). The LCBD values were regressed on the environmental variables to determine the factors that make LCBD vary along the river (adjusted $R^2 = 0.58$). Only two environmental variables were retained by backward elimination in regression: riverbed slope and BOD. Both variables had positive coefficients in the model, indicating that sites with high BD_{Total} either had a large slope (specially true at the headwaters) or were strongly eutrophic (high BOD). Note that regressing LCBD

values on environmental variables is not the same as canonical analysis of the community data. For the chord-transformed Doubs fish data, forward selection of environmental variables in RDA produced a different model (adjusted $R^2 = 0.61$) containing five significant variables at the 0.05 level: distance from the source, altitude, slope, dissolved oxygen, and BOD. The question in RDA is to identify the factors driving the observed variation in community composition; RDA truly analyses beta diversity by decomposing the total variance of the species data, i.e. BD_{Total} , into explained and residual components. By contrast, in regression analysis of the LCBD indices, the question is why some sites have higher degrees of uniqueness in species composition than others.

Four species contributed to beta diversity well above the mean of the 27 species: the stone loach (*Barbatula barbatula*, Balitoridae), the common bleak (*Alburnus alburnus*, Cyprinidae), the Eurasian minnow (*Phoxinus phoxinus*, Cyprinidae), and the brown trout (*Salmo trutta fario*, Salmonidae) which had the highest SCBD index. The chord-transformed abundances of these species varied the most among sites. The brown trout, Eurasian minnow and stone loach are found in the unpolluted sites with high LCBD upriver, which have high conservation status, whereas the common bleak is abundant in the eutrophic sites with agricultural pollution in the middle course of the river. Sites in the latter group, which also have high LCBD values, are in need of restoration.

One may wonder: For the coefficients that are appropriate for beta diversity studies, are the LCBD estimates similar or very different? Using the software in Appendix S4, calculation of LCBD was repeated for the 11 dissimilarities belonging to types II to IV, which are appropriate for beta assessment. The 11 LCBD vectors were quite similar: their mean Spearman correlation was 0.905. Kendall concordance analysis (Legendre 2005) showed that the contributions of all 11

vectors to the concordance of the group were significant. (These are not genuine tests of significance since the LCBD vectors were all computed from the same data; these results provide, however, a clustering validation criterion.) These results show that LCDB indices computed using all dissimilarities that were suitable for beta diversity assessment were highly concordant.

DISCUSSION

Different concepts of beta diversity

We will first address the appropriateness of using “beta diversity” to designate the approach described in this paper. We acknowledge that this is an unsettled issue. Authors, e.g. Anderson *et al.* (2011), have rightfully argued that there are several meanings and measures associated with the concept of beta diversity. Authors agree that alpha and beta diversities are essentially different; alpha measures how diversified the species are within a site, i.e. in a single row of the site-by-species data table \mathbf{Y} , whereas beta measures how diversified the sites are in species composition within a region, i.e. the variation among the rows of \mathbf{Y} . Some ecologists prefer to reserve the expression beta diversity for the additive or multiplicative approaches, and we will not dispute their choice.

However, if beta diversity can be seen as “the variation in species composition among sites”, as stated by many authors, then the variance of \mathbf{Y} , which specifically measures that variation, certainly qualifies as a measure of beta. The literature is growing that adopts this broader concept and measure of beta, because it links the ecological concept of beta diversity to methods of analysis that can be applied to test hypotheses about the mechanisms that generate and maintain beta diversity in ecosystems (subsection “Multiple ways of partitioning total beta diversity”). Those who prefer to limit the meaning of beta diversity to the additive or

multiplicative approaches do not deny that variation in species composition among sites can be analysed, and hypotheses tested, but they prefer to call that variation by some other name, e.g. compositional heterogeneity among sites. Compositional heterogeneity — be it called beta diversity, or not — measures community differentiation, which results from evolutionary and ecological processes operating at several spatial (from site to global) and temporal scales.

After proposing the concept in his seminal papers, Whittaker (1960, 1972) detailed different measures of beta diversity. One of his measures corresponds precisely to the variance of \mathbf{Y} measured through some dissimilarity coefficients, as will be shown in the next subsection. We are in good company here. Ecologists largely agree with Whittaker (1972) that beta diversity conceptually corresponds to *the variation in species composition among sites in the geographic region of interest*. [Whittaker used a slightly different expression, “the extent of differentiation of communities along habitat gradients”. He was interested in the response of communities to environmental variation, hence his interest for ordination methods.] Legendre *et al.* (2005) were perhaps the first to use precisely that expression, based on their reading of Whittaker, and they were followed in its use by many authors, including Anderson *et al.* (2006), Anderson *et al.* (2011) and Baselga and Orme (2012). Leaving the terminological issue aside, we may discuss what are the different ways of estimating the variation in species composition among sites, or beta diversity. Alternative estimation methods are not in opposition but complementary; each one offers a different way of explaining beta diversity, or expressing it in a way that makes it useful for ecological interpretation, impact assessment, or conservation studies.

Related approaches to beta diversity assessment

In this paper, we used the total variance of \mathbf{Y} as an estimate of beta diversity for a region of interest (eq. 3). Alternative equations have been proposed by Whittaker (1972), Ricotta &

485 Marignani (2007) and Anderson *et al.* (2006). We will now show that these proposals are
 486 special cases of eq. 3 or are related to it.

487 In section “Equivalent ways of computing $\text{Var}(\mathbf{Y})$ ”, we saw that $\text{SS}(\mathbf{Y})$ can be computed as
 488 the sum of the squared dissimilarities divided by n (eq. 8). This is appropriate for the Euclidean
 489 distance and for dissimilarities that have the property of being Euclidean (P13). Appendix S2
 490 shows that SS_{Total} can also be computed in that way for dissimilarities that do not lead to a fully
 491 Euclidean representation; these will not concern us in the present paragraph. Several
 492 dissimilarities, coded 1 for P13 in Table 2, are Euclidean only when taking their square roots; the
 493 square-rooted distances form matrix $\mathbf{D}^{(0.5)} = [D_{hi}^{0.5}]$. That group includes the Canberra metric,
 494 Whittaker’s index, the percentage difference (*alias* Bray-Curtis) and Wishart’s coefficient. Many
 495 of the incidence-based (i.e. binary) coefficients are also in that situation, including the widely
 496 used Jaccard, Sørensen and Ochiai coefficients (Legendre & Legendre 2012, Table 7.2). We will
 497 show here that the method of calculation of beta diversity proposed in other papers is equivalent
 498 to DB_{Total} of the present paper if $\mathbf{D}^{(0.5)}$ is used for the calculation.

499 (a) Whittaker (1972, p. 233) stated that “The mean CC [Jaccard or Sørensen coefficient of
 500 community] for samples of a set compared with one another in all possible directions is one
 501 expression [of] their relative dissimilarity, or beta differentiation”. The mean is obtained by
 502 summing the dissimilarities and dividing by the number of dissimilarities in the half-matrix,
 503 $n(n-1)/2$. This is equivalent to computing eqs 8 and 3 on the square-rooted dissimilarities
 504 (matrix $\mathbf{D}^{(0.5)}$) and multiplying by 2. Hence, Whittaker’s formula only differs by a factor 2 from
 505 DB_{Total} computed from $\mathbf{D}^{(0.5)}$.

506 (b) There is also a relationship between the equation for DB_{Total} used in this paper and the

suggestion of Ricotta & Marignani (2007) to estimate beta diversity by Rao's (1982) quadratic
 entropy, $Q = \sum_{h=1}^{n-1} \sum_{i=h+1}^n \delta_{hi} p_h p_i$, where p_i and p_h contain the relative abundance of sampling
 units i and h , respectively, and δ_{hi} is the dissimilarity between i and h computed with any measure
 of one's choice. If all sampling units are considered to be equally important, say $p_i = 1/n$,
 then $Q = \frac{1}{n^2} \sum_{h=1}^{n-1} \sum_{i=h+1}^n \delta_{hi}$, which is very close to DB_{Total} computed from $\mathbf{D}^{(0.5)}$ through eq. 8
 followed by eq. 3. The difference is that the last division is by n in Q instead of $(n-1)$ in eq. 3.
 (c) The beta diversity statistic developed by Anderson *et al.* (2006) belongs to the same family as
 DB_{Total} . It is the *sum of the dissimilarities from the sampling units to the group centroid* in
 multivariate space divided by n , producing a maximum likelihood estimate of the variance. It
 differs from DB_{Total} , which is the *sum of the squared dissimilarities from the sampling units to the*
group centroid divided by $(n-1)$ (eq. 3). The squared dissimilarities from the sampling units to
 the group centroid are found in vector $[SS_i]$ obtained by eqs 9 and 10a computed from \mathbf{D} .
 Because it can be computed from any dissimilarity matrix, the Anderson *et al.* (2006) statistic can
 be computed from \mathbf{D} or $\mathbf{D}^{(0.5)}$, both producing a different statistic than DB_{Total} .

Regarding the choice of a dissimilarity measure and the equivalence of the beta diversity
 approaches described in the last paragraphs, different situations should be considered. (a) For
 dissimilarity measures that are not Euclidean for \mathbf{D} but are Euclidean for $\mathbf{D}^{(0.5)}$, then the
 approaches of Whittaker (1972) and Ricotta & Marignani (2007) are essentially equivalent to the
 calculation of DB_{Total} in the present paper. (b) If the dissimilarity measure can be obtained by
 applying a transformation to the original data (Appendix S1) followed by the computation of the
 Euclidean distance, the equivalence between these methods holds in the transformed space and
 BD_{Total} can be computed by applying eqs 2 and 3 to the transformed data. (c) If the dissimilarity

measure cannot be obtained by applying a transformation to the original data followed by Euclidean distance calculation, the distances to the centroid can still be computed using the square root of eq. 10a. This result holds for non-Euclidean embeddable dissimilarities as well, although with some additional complexities (Anderson 2006; Appendix S2).

Multiple ways of partitioning total beta diversity

The strongest advantage of adopting the present approach to the analysis of beta diversity lies in the possibility of partitioning the total sum-of-squares of the community composition data into additive components. The total variance is the basic currency of many statistical methods, univariate and multivariate, through which $\text{Var}(\mathbf{Y})$ can be partitioned in different ways. Available partitioning methods include the following.

1. *Contributions of individual species.* — The SS_{Total} statistic can be partitioned into species contributions to beta diversity (SCBD_j , eq. 4b). SCBD indices can, in principle, be computed for raw or transformed abundance data, but it should in practice be limited to data subjected to the Hellinger or chord transformations, which are the only two that correspond to distances suitable for beta assessment. After centring, the SCBD values have signs which indicate the species that vary more [or less] than the mean across the sites. A mathematical limitation restrains the use of SCBD coefficients: they can only be computed from raw or transformed data tables with species in columns; they cannot be computed from a \mathbf{D} matrix. Calculating SCBD indices is useful to determine which species exhibit large variations across the study area. Note that SCBD indices do not have the same interpretation as indicator species for groups of sites (Dufrêne & Legendre 1997, De Cáceres & Legendre 2009). The sites where species with large SCBD values are abundant and dominate the community will normally also have large LCBD indices, as we found

in our example.

2. *Contributions of individual sampling units.* — Likewise, the SS_{Total} statistic can be partitioned into local contributions of individual sampling units to beta diversity ($LCBD_i$, eq. 5b or 10b). The $LCBD$ values, which can be mapped, indicate the sites that contribute more [or less] than the mean to beta diversity. $LCBD$ are comparative indicators of site uniqueness; hence, large $LCBD$ values indicate sites that have strongly different species compositions. In conservation biology, large $LCBD$ values indicate sites that have unusual species combinations and high conservation value, as well as degraded and species-poor sites in need of ecological restoration. $LCBD$ may be inversely correlated with species richness, as in our example, but in other ecosystems large $LCBD$ s may indicate rare species combinations that are worth studying in more detail.

In data analysis, sites with high $LCBD$ may be removed before simple or canonical ordination because they may have an undue influence on the results. This may prove a useful criterion to remove sites prior to ordination, instead of other criteria like low species richness.

3. *Within- and among-group contributions.* — Groups of sites may be known *a priori* from the sampling design, or they may be obtained by clustering based on the environmental variables. For these groups of sites, the total sum-of-squares of the species data can be divided by multivariate analysis of variance (computed using MANOVA or canonical analysis) into within- and among-group sums of squares. Alternatively, groups of sites where the species respond in the same way to environmental variables can be identified by multivariate regression tree analysis.

4. *Simple and canonical ordination.* — The total sum-of-squares, which estimates beta diversity, can be partitioned into orthogonal axes by simple ordination methods (PCA, CA, PCoA).

Alternatively, SS_{Total} can be partitioned by canonical analysis (RDA or CCA) into orthogonal

axes related to the environmental variables.

5. *Contributions of sets of explanatory factors*. — SS_{Total} can be partitioned as a function of different sets of explanatory variables by variation partitioning (Borcard *et al.* 1992; Peres-Neto *et al.* 2006). Partitioning can be done, for example, between different sets of environmental variables, or between explanatory matrices representing environmental and spatial variables (e.g. sets of spatial eigenfunctions), depending on the hypotheses under study. This is a major approach for estimating the relative contributions of groups of explanatory variables representing different hypotheses about the origin of beta diversity.

6. *Spatial scales*. — SS_{Total} can be partitioned as a function of spatial scales by spatial eigenfunction analysis. See Legendre & Legendre (2012, Chapter 14) for a review of these methods. These and other methods of multivariate multiscale analysis were also reviewed by Dray *et al.* (2012).

7. *Multivariate variogram and multiscale ordination*. — SS_{Total} can also be partitioned into spatial scales by multivariate variogram analysis (Wagner 2003). Furthermore, the species-environment relation, which represents a portion of SS_{Total} , can be partitioned into spatial scales by multiscale ordination; see Wagner (2003, 2004) and Legendre & Legendre (2012, Section 14.4).

Choosing a dissimilarity index for beta diversity assessment

Analysing the spatial variation in species composition necessarily implies choosing a dissimilarity coefficient, either implicitly or explicitly (Legendre *et al.* 2005, Anderson *et al.* 2006). Choosing an appropriate coefficient is crucial to ensure the interpretation of the results and allow the comparison of beta diversity estimates among regions and types of organisms.

In this paper, we studied several properties of coefficients, separating those that were purely mathematical from those that had an ecological interpretation. This conceptual separation was important to help users make choices on ecological grounds. Comparison of the 16 selected dissimilarity coefficients based on 14 ecological, statistical and mathematical properties led to a model where the coefficients were divided into five main types. Three of those types are suitable for beta diversity studies and comparison of beta diversity estimates computed from different ecological data sets. These different types of coefficients can be used to address different questions.

Among the unsuitable coefficients are the Manhattan and Euclidean distances. As shown in this paper, these distances are appropriate for beta diversity assessments only after transformation of the raw abundance data. In the case of the Manhattan distance (L1 norm), the natural transformation is the division of each value by the total abundance, which leads to the Whittaker coefficient. In the case of the Euclidean distance (L2 norm), the natural transformation is the division of each value by the norm of the row vector, which leads to the chord distance. The Hellinger distance is the chord distance computed on square-root-transformed abundance data.

When choosing a coefficient, users should check the properties the coefficient has, and determine whether they are suitable for the objectives of the study. Further research is needed about the mathematical and ecological properties of dissimilarity coefficients and the situations where these properties are desirable or needed.

ACKNOWLEDGEMENTS

This paper is dedicated to Dr. Francesc Oliva, who fostered the interest of M. De Cáceres for dissimilarity coefficients and their use in ecology. Our thanks to Daniel Borcard who provided

comments on a first draft of the manuscript, and to Anne Chao and two other anonymous referees who provided very interesting comments that helped us improve the paper. This research was supported by a NSERC grant no. 7738 to P. Legendre. M. De Cáceres was supported by research projects BIONOVEL (CGL2011-29539/BOS) and MONTES (CSD2008-00040) funded by the Spanish Ministry of Education and Science.

REFERENCES

1.
Anderson, M.J. (2006). Distance-based tests for homogeneity of multivariate dispersions. *Biometrics*, 62, 245–253.
2.
Anderson, M.J., Ellingsen, K.E. & McArdle, B.H. (2006). Multivariate dispersion as a measure of beta diversity. *Ecol. Lett.*, 9, 683–693.
3.
Anderson, M.J., Crist, T.O., Chase, J.M., Vellend, M., Inouye, B.D., Freestone, A.L. *et al.* (2011). Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. *Ecol. Lett.*, 14, 19–28.
4.
Baselga, A. (2010). Partitioning the turnover and nestedness components of beta diversity. *Global Ecol. Biogeogr.*, 19, 134–143.
5.
Baselga, A. & Orme, C.D.L. (2012). betapart: an R package for the study of beta diversity. *Methods Ecol. Evol.*, 3: 808–812.
- 6.

- 640 Bloom, S.A. (1981). Similarity indices in community studies: potential pitfalls. *Mar. Ecol.*
641 *Progr. Ser.*, 5, 125–128.
- 642 7.
- 643 Borcard, D., Gillet, F. & Legendre, P. (2011). *Numerical ecology with R*. Use R! series, Springer
644 Science, New York.
- 645 8.
- 646 Borcard, D., Legendre, P. & Drapeau, P. (1992). Partialling out the spatial component of
647 ecological variation. *Ecology*, 73, 1045–1055.
- 648 9.
- 649 Bray, R.J. & Curtis, J.T. (1957). An ordination of the upland forest communities of southern
650 Wisconsin. *Ecol. Monogr.*, 27, 325–349.
- 651 10.
- 652 Cardoso, P., Borges, P.A.V. & Veech, J.A. (2009). Testing the performance of beta diversity
653 measures based on incidence data: the robustness to undersampling. *Divers. Distrib.*, 15,
654 1081–1090.
- 655 11.
- 656 Chao, A., Chazdon, R.L., Colwell, R.K. & Shen, T.J. (2006). Abundance-based similarity indices
657 and their estimation when there are unseen species in samples. *Biometrics* 62, 361–371.
- 658 12.
- 659 Chao, A., Chiu, C.-H. & Hsieh, T.C. (2012). Proposing a resolution to debates on diversity
660 partitioning. *Ecology*, 93, 2037–2051.
- 661 13.
- 662 Clark, P.J. (1952). An extension of the coefficient of divergence for use with multiple characters.
663 *Copeia*, 1952, 61–64.

- 664 14.
- 665 Clarke, K.R., Somerfield, P.J. & Chapman, M.G. (2006). On resemblance measures for
666 ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray–Curtis
667 coefficient for denuded assemblages. *J. Exp. Mar. Biol. Ecol.*, 330, 55–80.
- 668 15.
- 669 Czekanowski, J. (1909). Zur Differentialdiagnose der Neandertalgruppe. *Korrespondenz-Blatt*
670 *deutsch. Ges. Anthropol. Ethnol. Urgesch.*, 40, 44–47.
- 671 16.
- 672 De Cáceres, M. and Legendre, P. (2009). Associations between species and groups of sites:
673 indices and statistical inference. *Ecology*, 90, 3566–3574.
- 674 17.
- 675 De Cáceres, M., Legendre, P., Valencia, R., Cao, M., Chang, L.-W., Chuyong, G. *et al.* (2012).
676 The variation of tree beta diversity across a global network of forest plots. *Global Ecol.*
677 *Biogeogr.*, 21, 1191–1202.
- 678 18.
- 679 Dray, S., Péliissier, R., Couteron, P., Fortin, M.-J., Legendre, P., Peres-Neto, P.R. *et al.* (2012).
680 Community ecology in the age of multivariate multiscale spatial analysis. *Ecol. Monogr.*,
681 82, 257–275.
- 682 19.
- 683 Dufrêne, M. & Legendre P. (1997). Species assemblages and indicator species: the need for a
684 flexible asymmetrical approach. *Ecol. Monogr.*, 67, 345–366.
- 685 20.
- 686 Ellison, A.M. (2010). Partitioning diversity. *Ecology*, 91, 1962–1963.

- 687 21.
- 688 Faith, D.P., Minchin, P.R. & Belbin, L. (1987). Compositional dissimilarity as a robust measure
689 of ecological distance. *Vegetatio*, 69, 57–68.
- 690 22.
- 691 Gower, J.C. (1966). Some distance properties of latent root and vector methods used in
692 multivariate analysis. *Biometrika*, 53, 325–338.
- 693 23.
- 694 Gower, J.C. & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients.
695 *J. Classif.*, 3, 5–48.
- 696 24.
- 697 Hajdu, L.J. (1981). Graphical comparison of resemblance measures in phytosociology. *Vegetatio*,
698 48, 47–59.
- 699 25.
- 700 Hubálek, Z. (1982). Coefficients of association and similarity, based on binary (presence-
701 absence) data: an evaluation. *Biol. Rev.*, 57, 669–689.
- 702 26.
- 703 Hurlbert, S.H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecol.*
704 *Monogr.*, 54, 187–211.
- 705 27.
- 706 Jaccard, P. (1900). Contribution au problème de l’immigration post-glaciaire de la flore alpine.
707 *Bull Soc. Vaudoise Sci. Nat.*, 36, 87–130.
- 708 28.
- 709 Janson, S. & Vegelius, J. (1981). Measures of ecological association. *Oecologia*, 49, 371–376.

- 710 29.
- 711 Janssen, J.G.M. (1975). A simple clustering procedure for preliminary classification of very large
712 sets of phytosociological relevés. *Vegetatio*, 30, 67–71.
- 713 30.
- 714 Jost, L. (2007). Partitioning diversity into independent alpha and beta components. *Ecology*, 88,
715 2427–2439.
- 716 31.
- 717 Koleff, P., Gaston, K.J. & Lennon, J.J. (2003). Measuring beta diversity for presence-absence
718 data. *J. Anim. Ecol.*, 72, 367–382.
- 719 32.
- 720 Kraft, N.J.B., Comita, L.S., Chase, J.M., Sanders, N.J., Swenson, N.G., Crist, T.O. *et al.* (2011).
721 Disentangling the drivers of diversity along latitudinal and elevational gradients. *Science*,
722 333, 1755–1758.
- 723 33.
- 724 Kulczynski, S. (1928). Die Pflanzenassoziationen der Pieninen. *Bull. Int. Acad. Pol. Sci. Lett. Cl.*
725 *Sci. Math. Nat. Ser. B*, Suppl. II (1927), 57–203.
- 726 34.
- 727 Lance, G.N. & Willams, W.T. (1967). Mixed-data classificatory programs. I. Agglomerative
728 systems. *Aust. Comput. J.*, 1, 15–20.
- 729 35.
- 730 Lebart, L. & Fénelon, J.P. (1971). *Statistique et informatique appliquées*. Dunod, Paris, France.
- 731 36.
- 732 Legendre, P. (2005). Species associations: the Kendall coefficient of concordance revisited. *J.*
733 *Agr. Biol. Envir. S.*, 10, 226–245.

- 734 37.
- 735 Legendre, P. & Anderson, M.J. (1999). Distance-based redundancy analysis: testing multispecies
 736 responses in multifactorial ecological experiments. *Ecol. Monogr.*, 69, 1–24.
- 737 38.
- 738 Legendre, P., Borcard, D. & Peres-Neto, P.R. (2005). Analyzing beta diversity: partitioning the
 739 spatial variation of community composition data. *Ecol. Monogr.*, 75, 435–450.
- 740 39.
- 741 Legendre, P. & Fortin, M.-J. (2010). Comparison of the Mantel test and alternative approaches
 742 for detecting complex multivariate relationships in the spatial analysis of genetic data.
 743 *Mol. Ecol. Resour.*, 10, 831–844.
- 744 40.
- 745 Legendre, P. & Gallagher, E.D. (2001). Ecologically meaningful transformations for ordination
 746 of species data. *Oecologia*, 129, 271–280.
- 747 41.
- 748 Legendre, P. & Legendre, L. (2012). *Numerical ecology*. 3rd English edition. Elsevier Science
 749 BV, Amsterdam.
- 750 42.
- 751 Nekola, J.C. & White, P.S. 1999). The distance decay of similarity in biogeography and ecology.
 752 *J. Biogeogr.*, 26, 867–878.
- 753 43.
- 754 Odum, E.P. (1950). Bird populations of the Highlands (North Carolina) Plateau in relation to
 755 plant succession and avian invasion. *Ecology*, 31, 587–605.
- 756 44.
- 757 Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O’Hara, R.B. *et al.* (2012).

- 758 *vegan: Community ecology package. R package version 2.0-3. Available at:*
 759 <http://cran.r-project.org/web/packages/vegan/>.
 760 45.
- 761 Orlóci, L. (1967). An agglomerative method for classification of plant communities. *J. Ecol.*, 55,
 762 193–206.
- 763 46.
- 764 Orlóci, L. (1978). *Multivariate analysis in vegetation research*. 2nd edition. Dr. W. Junk B. V.,
 765 The Hague, The Netherlands.
- 766 47.
- 767 Pelissier, R., Couteron, P., Dray, S. & Sabatier, D. (2003). Consistency between ordination
 768 techniques and diversity measurements: two strategies for species occurrence data.
 769 *Ecology*, 84, 242–251.
- 770 48.
- 771 Peres-Neto, P.R., Legendre, P., Dray, S. & Borcard, D. (2006). Variation partitioning of species
 772 data matrices: estimation and comparison of fractions. *Ecology*, 87, 2614–2625.
- 773 49.
- 774 Rao, C.R. (1982). Diversity and dissimilarity coefficients: a unified approach. *Theor. Popul.*
 775 *Biol.*, 21, 24–43.
- 776 50.
- 777 Rao, C.R. (1995). A review of canonical coordinates and an alternative to correspondence
 778 analysis using Hellinger distance. *Qüestió (Quaderns d'Estadística i Investivació*
 779 *Operativa)*, 19, 23–63.
- 780 51.
- 781 Ricotta, C. & Marignani, M. (2007). Computing B-diversity with Rao's quadratic entropy: a

- 782 change of perspective. *Divers. Distrib.*, 13, 237–241.
- 783 52.
- 784 Stephenson, W., Williams, W.T. & Cook, S.D. (1972). Computer analyses of Petersen's original
- 785 data on bottom communities. *Ecol. Monogr.*, 42, 387–415.
- 786 53.
- 787 Vellend, M. (2001). Do commonly used indices of beta-diversity measure species turnover? *J.*
- 788 *Veg. Sci.*, 12, 545–552.
- 789 54.
- 790 Verneaux, J. (1973). Cours d'eau de Franche-Comté (Massif du Jura). Recherches écologiques
- 791 sur le réseau hydrographique du Doubs – Essai de biotypologie. *Annales Scientifiques de*
- 792 *l'Université de Franche-Comté, Biologie Animale*, 3, 1–260.
- 793 55.
- 794 Wagner, H.H. (2003). Spatial covariance in plant communities: integrating ordination, variogram
- 795 modeling, and variance testing. *Ecology*, 84, 1045–1057.
- 796 56.
- 797 Wagner, H.H. (2004). Direct multi-scale ordination with canonical correspondence analysis.
- 798 *Ecology*, 85, 342–351.
- 799 57.
- 800 Whittaker, R.H. (1952). A study of summer foliage insect communities in the Great Smoky
- 801 Mountains. *Ecol. Monogr.*, 22, 1–44.
- 802 58.
- 803 Whittaker, R.H. (1960). Vegetation of the Siskiyou mountains, Oregon and California. *Ecol.*
- 804 *Monogr.*, 30, 279–338.
- 805 59.

806 Whittaker, R.H. (1972). Evolution and measurement of species diversity. *Taxon*, 21, 213–251.

807 60.

808 Wilson, M.V. & Shmida, A. (1984). Measuring beta diversity with presence-absence data. *J.*

809 *Ecol.*, 72, 1055–1064.

810 61.

811 Wishart, D. (1969). *CLUSTAN 1a user manual*. Computing Laboratory, University of St.

812 Andrews, St. Andrews, Fife, Scotland.

813 **SUPPORTING INFORMATION**

814 Additional Supporting Information may be found in the online version of this article:

815 **Appendix S1** Community composition data transformations.

816 **Appendix S2** Computing the total sum of squares from a dissimilarity matrix.

817 **Appendix S3** Details about the properties of dissimilarity coefficients.

818 **Appendix S4** The R function `beta.div()` computes estimates of total beta diversity as the total

819 variance in a community data table **Y**, as well as the derived SCBD and LCBD statistics, for 16

820 dissimilarity coefficients or the raw data table.

Table 1 Dissimilarity coefficients compared in this paper.

Dissimilarity	Abundance-based	Incidence-based	References	Coefficient no. in L&L ¹
Euclidean distance	$\sqrt{\sum_{j=1}^p [y_{1j} - y_{2j}]^2}$	$\sqrt{p \left(\frac{b+c}{a+b+c+d} \right)} = \sqrt{b+c}$		D ₁
Manhattan distance	$\sum_{j=1}^p y_{1j} - y_{2j} $	$p \left(\frac{b+c}{a+b+c+d} \right) = b+c$		D ₇
Modified mean character difference	$\frac{1}{pp} \sum_{j=1}^p y_{1j} - y_{2j} $	$\frac{b+c}{a+b+c}$	Legendre & Legendre (2012)	D ₁₉
Species profile distance	$\sqrt{\sum_{j=1}^p \left[\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right]^2}$	$\sqrt{\frac{b+c}{(a+b)(a+c)}}$	Legendre & Gallagher (2001)	D ₁₈
Hellinger distance	$\sqrt{\sum_{j=1}^p \left[\sqrt{\frac{y_{1j}}{y_{1+}}} - \sqrt{\frac{y_{2j}}{y_{2+}}} \right]^2}$	$\sqrt{2 \left(1 - \frac{a}{\sqrt{(a+b)(a+c)}} \right)}$	Rao (1995)	D ₁₇
Chord distance	$\sqrt{\sum_{j=1}^p \left[\frac{y_{1j}}{\sqrt{\sum_{k=1}^p y_{1k}^2}} - \frac{y_{2j}}{\sqrt{\sum_{k=1}^p y_{2k}^2}} \right]^2}$	$\sqrt{2 \left(1 - \frac{a}{\sqrt{(a+b)(a+c)}} \right)}$	Orlóci (1967)	D ₃
Chi-square distance	$\sqrt{y_{++} \sum_{j=1}^p \frac{1}{y_{+j}} \left[\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right]^2}$	NA ²	Lebart & Fénelon (1971)	D ₁₆
Coefficient of divergence	$\sqrt{\frac{1}{pp} \sum_{j=1}^p \left(\frac{y_{1j} - y_{2j}}{y_{1j} + y_{2j}} \right)^2}$	$\sqrt{\frac{b+c}{a+b+c}}$	Clark (1952)	D ₁₁
Canberra metric ³	$\frac{1}{pp} \sum_{j=1}^p \frac{ y_{1j} - y_{2j} }{(y_{1j} + y_{2j})}$	$\frac{b+c}{a+b+c}$	Lance & Willams (1967), Stephenson <i>et al.</i> (1972) for 1/pp	D ₁₀

Whittaker's index of association	$\frac{1}{2} \sum_{j=1}^p \left \frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right $	$\frac{1}{2} \left(\frac{b}{a+b} + \frac{c}{a+c} + \left \frac{a}{a+b} - \frac{a}{a+c} \right \right)$	Whittaker (1952)	D ₉
Percentage difference (<i>alias</i> Bray-Curtis dissimilarity ⁴)	$\frac{\sum_{j=1}^p y_{1j} - y_{2j} }{y_{1+} + y_{2+}}$	$\frac{b+c}{2a+b+c}$	Odum (1950)	D ₁₄
Wishart coefficient = (1 – similarity ratio)	$1 - \left[\frac{\sum_{j=1}^p y_{1j} y_{2j}}{\sum_{j=1}^p y_{1j}^2 + \sum_{j=1}^p y_{2j}^2 - \sum_{j=1}^p y_{1j} y_{2j}} \right]$	$\frac{b+c}{a+b+c}$	Wishart (1969), Janssen (1975)	
D = (1–Kulczynski coefficient)	$1 - \frac{1}{2} \left[\frac{\sum_{j=1}^p \min(y_{1j}, y_{2j})}{y_{1+}} + \frac{\sum_{j=1}^p \min(y_{1j}, y_{2j})}{y_{2+}} \right]$	$1 - \frac{1}{2} \left(\frac{a}{a+b} + \frac{a}{a+c} \right)$	Kulczynski (1928)	1 – S ₁₈
Abundance-based Jaccard ⁵	$\left(1 - \frac{UV}{U+V-UV} \right)$	$\frac{b+c}{a+b+c}$	Chao <i>et al.</i> (2006)	
Abundance-based Sørensen ⁵	$\left(1 - \frac{2UV}{U+V} \right)$	$\frac{b+c}{2a+b+c}$	Chao <i>et al.</i> (2006)	
Abundance-based Ochiai ⁵	$(1 - \sqrt{UV})$	$\left(1 - \frac{a}{\sqrt{(a+b)(a+c)}} \right)$	Chao <i>et al.</i> (2006)	

¹L&L = Legendre & Legendre (2012).

²NA: No binary form with parameters a , b and c for this coefficient, although it can be computed for presence-absence data.

³Division by pp (number of species excluding double zeros) introduced by Stephenson *et al.* (1972) and adopted by Oksanen *et al.* (2012).

⁴Coefficient first described by Steinhaus in the 1940's, then by Odum (1950) as the *percentage difference*. The Bray & Curtis (1957) paper described a new ordination method; the index described and used by these authors was Whittaker's dissimilarity, not the percentage difference which is more general. It is incorrect to attribute this coefficient to these authors.

⁵ U and V notation: see Chao *et al.* (2006).

Table 2 Properties P4 to P14 of the coefficients in Table 1. P1 to P3 (not shown) are fulfilled by all coefficients. Property descriptions are found in Appendix S3. 1 indicates that a coefficient has the property, 0 that it does not. For P13, code 2 indicates that both \mathbf{D} and $\mathbf{D}^{(0.5)}$ are Euclidean, 1 that only $\mathbf{D}^{(0.5)} = [D_{hi}^{0.5}]$ is Euclidean, and 0 that neither \mathbf{D} nor $\mathbf{D}^{(0.5)}$ are Euclidean. NA: there is no binary form for the chi-square distance, hence P10 could not be assessed. Last column: maximum possible dissimilarity value (D_{\max}) when it exists. P1 to P9 are essential properties for beta assessment; P10 to P14 describe additional properties, useful for special applications.

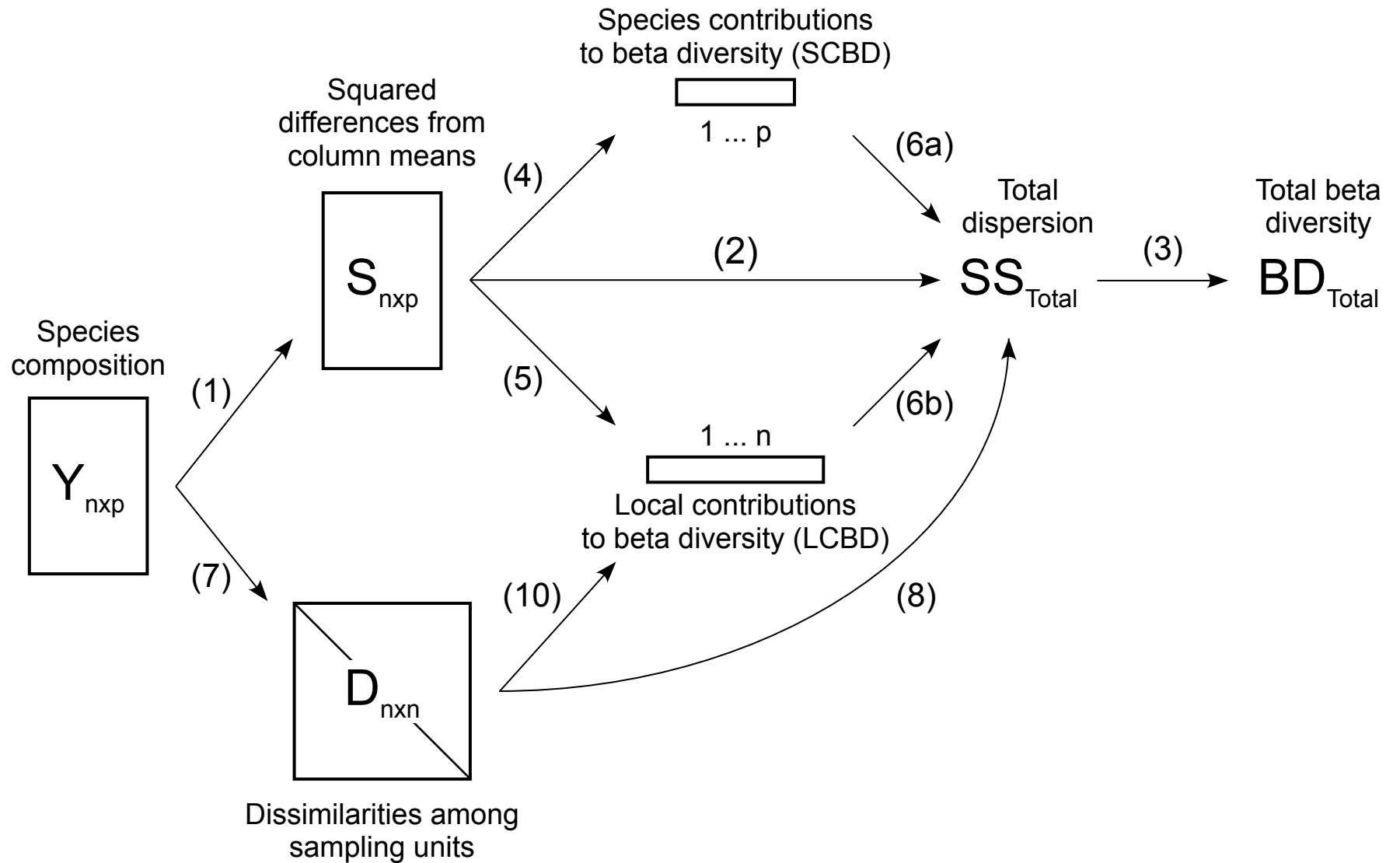
Dissimilarity	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	D_{\max}
Euclidean distance	0	0	1	0	0	0	0	0	0	2	1	—
Manhattan distance	0	0	1	0	0	0	0	0	0	1	0	—
Modified mean character difference	1	0	1	1	0	0	1	0	0	0	0	—
Species profile distance	1	0	0	0	1	1	0	1	0	2	1	$\sqrt{2}$
Hellinger distance	1	1	1	1	1	1	1	1	0	2	1	$\sqrt{2}$
Chord distance	1	1	1	1	1	1	1	1	0	2	1	$\sqrt{2}$
Chi-square distance	1	0	1	1	1	1	NA	0	0	2	1	$\sqrt{2y_{++}}$
Coefficient of divergence	1	1	1	1	1	1	1	0	0	2	0	1

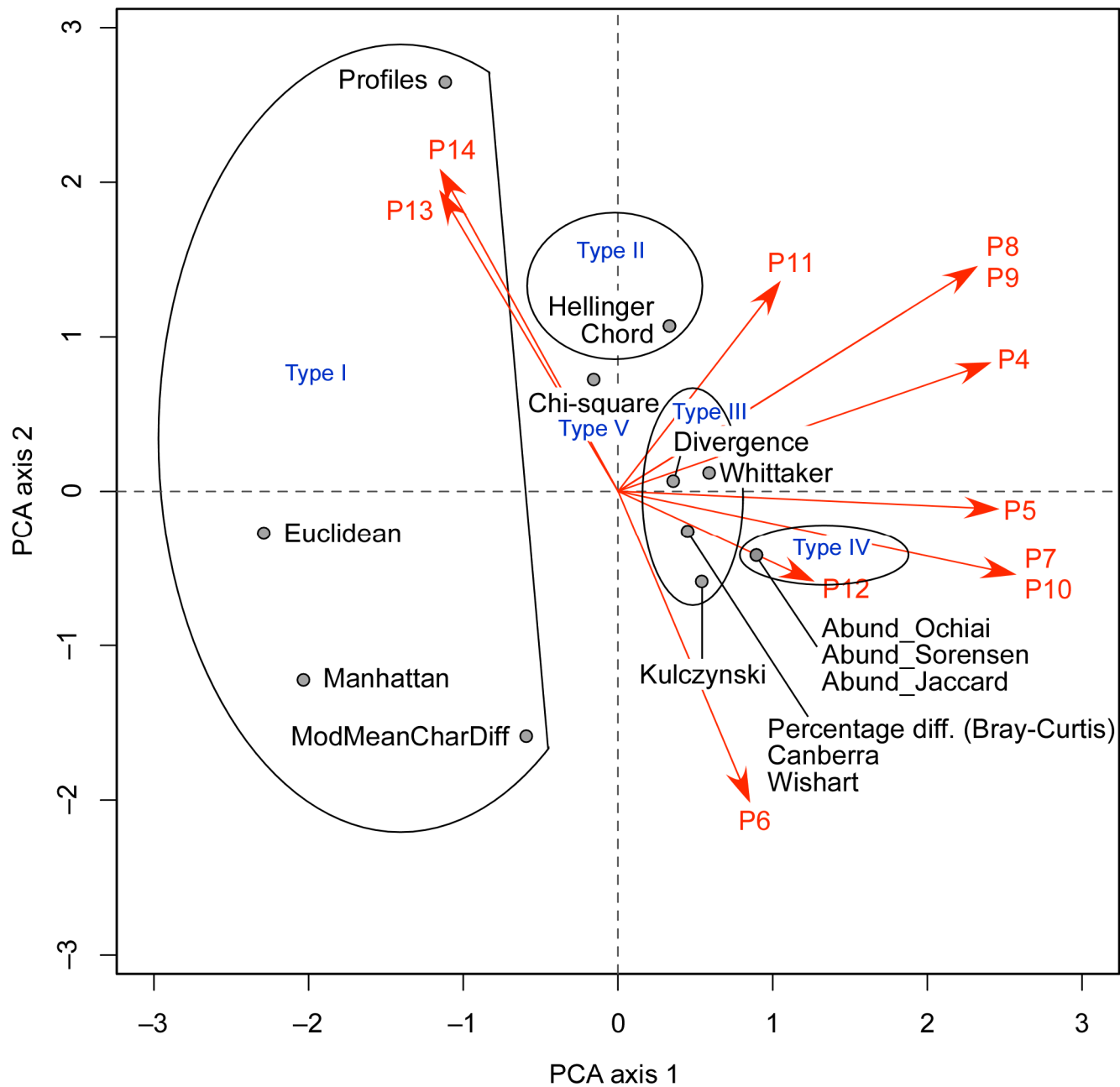
Figure 1 Schematic diagram representing the different ways of computing beta diversity as the total variance in the species composition data table \mathbf{Y} , as well as the contributions of individual species and sampling units. Numbers in parentheses refer to equations in the text.

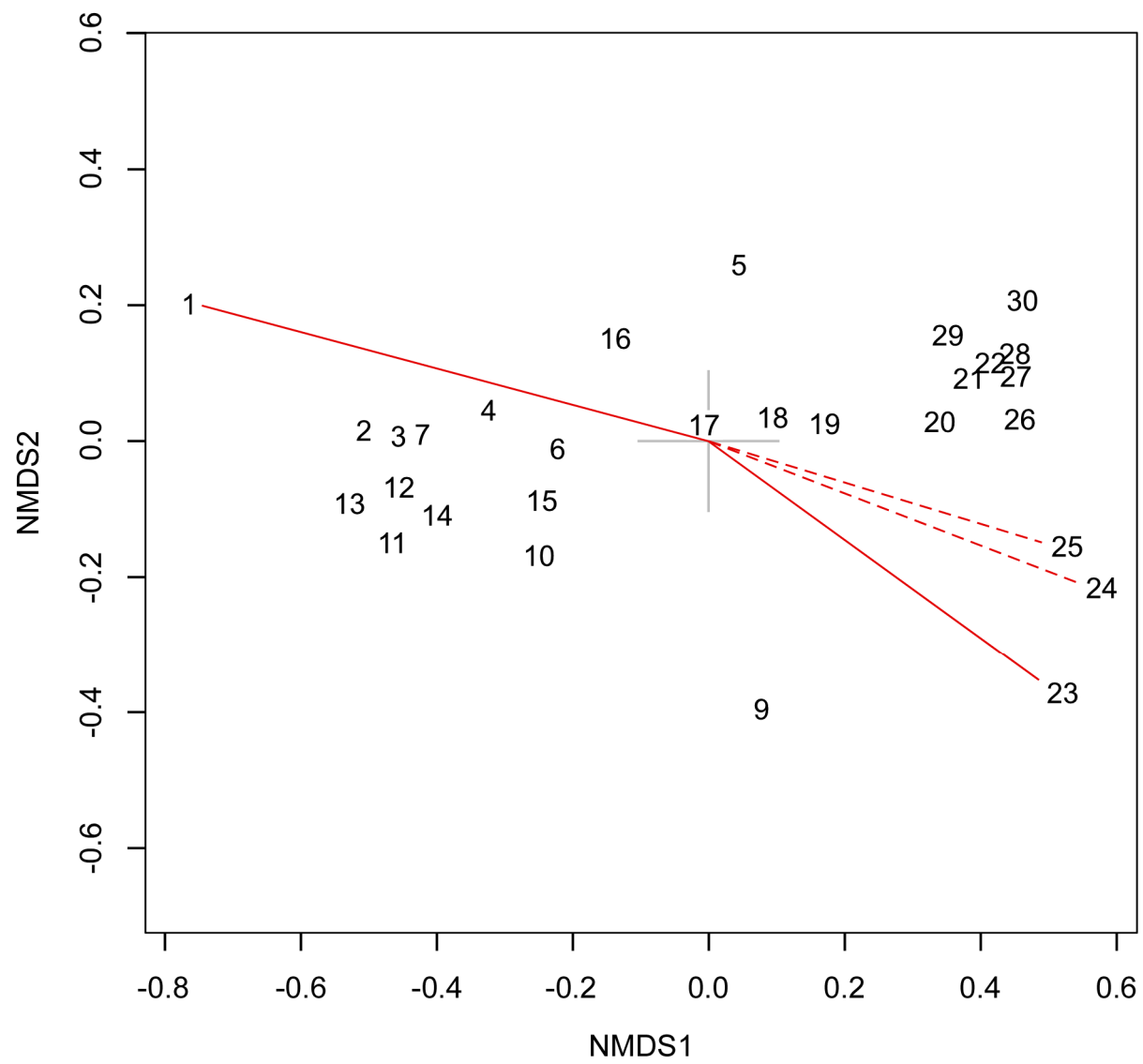
Figure 2 Principal component biplot relating properties P4 to P14 (red arrows) to the dissimilarity coefficients (gray points; see Table 1 for the full coefficient names). The five types of coefficients (blue labels), shown in the figure, are described in the text. PCA axis 1 accounts for 46% of the multivariate variation and axis 2 for 23%.

Figure 3 Ordination diagram of Doubs River fish data sites (nonmetric multidimensional scaling, nMDS; chord distance). SS_i indices are the squares of the distances of the sites to the multivariate centroid. The significant indices ($p < 0.05$) are represented by red lines joining the points to the centroid (full lines: $p < 0.05$ after Holm correction for 29 simultaneous tests).

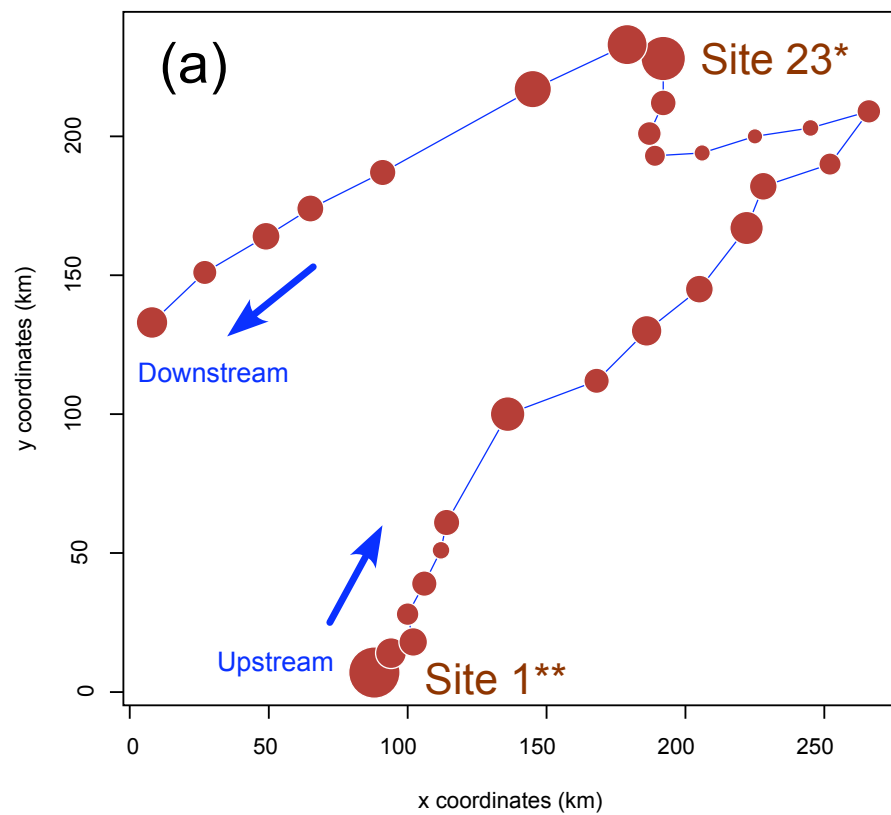
Figure 4 Maps of Doubs River (blue line) showing (a) the local contributions to beta diversity (LCBD) of the fish assemblage data and (b) the species richness at the 29 study sites. Size of the circles is proportional to the LCBD or richness values. Two sites have significant LCDB (or SS_i) indices at the 0.05 significance level after Holm correction for multiple testing: site 1 ($p = 0.003$) and site 23 ($p = 0.042$). The arrows indicate flow direction.







Map of LCBD



Map of Species Richness

