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

Pierre Legendre[1*], Miquel De Cáceres[2]

[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.*

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: 7491 # Text boxes: 0
References: 59 # 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 a community data matrix **Y** can be used as an estimate of beta diversity. We show how the total variance of **Y** can be calculated either directly or through a dissimilarity matrix. This measure can be generalized to any community dissimilarity index. We address the question of which index to use by coding 17 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, four of which are appropriate for beta diversity assessment. The total variance of **Y** links beta diversity with the analysis of community data by commonly used methods like ordination and ANOVA. Total beta can be partitioned in different ways: one can compute the contributions of individual species and sites to beta; local contributions to beta diversity (LCBD) are comparative indicators of the degree of ecological uniqueness of the sites under study. Moreover, total beta can be split 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.

Beta diversity is conceptually the variation in species composition among sites within a geographic area of interest (Whittaker 1960). 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 a non-directional approach to the study of community variation through space; it does not refer to any explicit gradient but simply focuses on the variation in community composition among the sampling units. Both approaches are legitimate.

Regardless of whether beta diversity is defined as directional or non-directional, one can be interested in summarizing it by a single number. 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. 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) and Koleff *et al.* (2003) 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 concluded that it would be most useful (he wrote: "a real breakthrough") to have a method to estimate beta diversity without calling upon alpha and gamma (computational independence, which does not imply statistical independence). Such an approach exists: the total variance in the community data table **Y** is a single-number estimate of beta diversity (Pelissier *et al.* 2003, Legendre *et al.* 2005, Anderson *et al.* 2006). It is computed without reference to the values of alpha and gamma and its statistical dependence on gamma can be removed (Kraft *et al*. 2011, De Cáceres *et al*. 2012). Most importantly, it allows ecologists to 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.

This paper focuses on exploring the advantages of estimating beta diversity as the total variation of the community data **Y** (BD_{Total}). (1) In a first section, we will 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. The second method is appealing because it derives the beta estimate using a dissimilarity function designed for the analysis of community data. (2) There are, however, many different dissimilarity coefficients, and users

are faced with the problem of choosing from among them. A detailed analysis of 17 dissimilarity coefficients will be undertaken in the following sections. (3) We will then present an example to illustrate the calculation of beta as the total variance in **Y** and the contributions of individual species and sampling units. (4) We will 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 will show that the total variance of **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 VARIANCE OF THE SPECIES DATA

Equivalent ways of computing Var(Y)

This section demonstrates that there are two equivalent ways of computing the total variance of community composition data **Y**. The first one is straightforward, it is simply the variance of **Y**. The second one is based upon the dissimilarity measures developed by ecologists during more than a century of field surveys. It 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 $\mathbf{y}_1$, $\mathbf{y}_2$, … $\mathbf{y}_p$ of **Y**) observed in n sampling units (row vectors $\mathbf{x}_1$, $\mathbf{x}_2$, … $\mathbf{x}_n$ of **Y**). We will use indices i and h for sampling units, index j for species, and y_{ij} for individual values in **Y**. The total variance of **Y**, noted Var(**Y**), can be computed as follows:

1. Sums of squares. — The usual way to obtain Var(**Y**) consists in computing a matrix of squared deviations from the column means. Let **S** be a $n \times p$ rectangular matrix where each element s_{ij} contains the square of the difference between the y_{ij} value and the mean value of the corresponding jth species:

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

s_{ij} is zero if all sites have the same abundance for species j. If we sum all values of **S**, we obtain the total sum of squares of the species composition data:

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

This quantity forms the basis of $\mathrm{BD}_{\mathrm{Total}}$, which is the index of beta diversity whose properties are developed in this section:

$$\mathrm{BD}_{\mathrm{Total}} = \mathrm{Var}(\mathbf{Y}) = \mathrm{SS}_{\mathrm{Total}} / (n-1) \; . \qquad (3)$$

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. $\mathrm{SS}_{\mathrm{Total}}$ and $\mathrm{Var}(\mathbf{Y}) = \mathrm{BD}_{\mathrm{Total}}$ 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, although there is a clear advantage in using Var(**Y**) for comparisons among regions.

An advantage of conceiving beta as the total variation in **Y** is that $\mathrm{SS}_{\mathrm{Total}}$ allows the assessment of the *contributions of individual species* or *individual sampling units to the overall beta diversity*. That is, one can compute the sum of squares corresponding to the jth species,

$$\mathrm{SS}_j = \sum_{i=1}^{n} s_{ij} \qquad (4a)$$

which is the contribution of species *j* to the overall beta diversity. The *relative* contribution of species *j* to beta, which we can call *Species Contribution to Beta diversity* (SCBD), is thus:

$$\mathrm{SCBD}_j = \mathrm{SS}_j / \mathrm{SS}_{\mathrm{Total}} \quad . \qquad (4b)$$

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

$$\mathrm{SS}_i = \sum_{j=1}^{p} s_{ij} \quad . \qquad (5a)$$

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. The SS_i values thus represent a genuine partitioning of beta diversity among the sites. 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 can call *Local Contribution to Beta diversity* (LCBD_i), is thus:

$$\mathrm{LCBD}_i = \mathrm{SS}_i / \mathrm{SS}_{\mathrm{Total}} \quad . \qquad (5b)$$

LCBD values can be mapped, as in the ecological illustration below. They represent *the degree of uniqueness of the sampling units in terms of community composition*. Mapping the centred values using different symbols would highlight the sites with LCBD values higher and lower than the mean.

Hence, the two decompositions of $\mathrm{SS}_{\mathrm{Total}}$ are:

$$\mathrm{SS}_{\mathrm{Total}} = \sum_{j=1}^{p} \mathrm{SS}_j \quad \text{and} \quad \mathrm{SS}_{\mathrm{Total}} = \sum_{i=1}^{n} \mathrm{SS}_i \, . \qquad (6a,b)$$

2. Dissimilarity. — There is an alternative path starting from **Y** and leading to $\mathrm{SS}_{\mathrm{Total}}$ (Fig. 1). That is, $\mathrm{SS}_{\mathrm{Total}}$ 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):

$$\mathrm{SS}_{\mathrm{Total}} = \frac{1}{n} \sum_{h=1}^{n-1} \sum_{i=h+1}^{n} D_{hi}^2 \quad (8)$$

That is, one can obtain $\mathrm{SS}_{\mathrm{Total}}$ by summing the squared distances in the upper or lower half of matrix **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). This statistic will be generalized below to other dissimilarity indices, which may or may not have the Euclidean property (P13 below). Working with matrix **D** instead of the matrix of squared centred values **S** entails the drawback of loosing track of the species. Because **D** is computed among sampling units over all species, the contributions of individual species cannot be recovered from **D**.

It is still possible, however, to calculate the contribution of individual sampling units from **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 **D**. In PCoA, prior to eigen-decomposition, the distance matrix is transformed into matrix $\mathbf{A} = [a_{hi}] = -0.5 D_{hi}^2$, then centred using the equation

$$\boldsymbol{\Delta}_1 = \left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n}\right) \mathbf{A} \left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n}\right) \quad (9)$$

where **I** is an identity matrix of size n, **1** is a vector of ones (length n) and **1'** is its transpose (Legendre & Legendre 2012, eqs 9.40 and 9.42). The diagonal elements of matrix $\boldsymbol{\Delta}_1$ are the SS_i values, or the squared distances of the points to the multivariate centroid of **Y**, and also to the centroid of the principal coordinate space:

$$[SS_i] = \text{diag}(\boldsymbol{\Delta}_1) \quad (10a)$$

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

$$[LCBD_i] = \text{diag}(\boldsymbol{\Delta}_1)/SS_{Total} \quad (10b)$$

To summarize:

• One can compute the total sum of squares in the community data **Y**, SS_{Total}, from either the community composition matrix **Y** (eq. 2) or from a Euclidean distance matrix **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, $BD_{Total} = \text{Var}(\mathbf{Y})$ (eq. 3).

• The contribution of the *i*th sampling unit to overall beta diversity can be computed using eq. 5a. The SS_i values are also found in the diagonal elements of matrix $\boldsymbol{\Delta}_1$ (eqs 9 and 10a). The relative contributions are computed using eqs 5b and 10b.

• The contribution of species *j* to the overall beta diversity, SS_j, is computed using eq. 4a, and the relative contributions are computed using eq. 4b.

Individual contributions of species and sampling units are useful complements to BD_{Total}. They are new and valuable tools for the assessment and interpretation of beta diversity, as will be shown in the illustrative example.

From the Euclidean distance to ecological dissimilarities

The Euclidean distance is known to be inappropriate for the analysis of community composition data sampled under varying environmental conditions (Orlóci 1978; Legendre & Gallagher 2001) because such data contain many double zeros, which should be analysed using double-zero asymmetrical coefficients (property P3 in "Properties of dissimilarity coefficients"). The Euclidean distance is not double-zero asymmetrical (*sensu* Legendre &

Legendre 2012, Section 7.4.1) and is thus inappropriate in most instances. An appropriate method consists in computing a dissimilarity matrix **D** using a chosen coefficient, instead of the Euclidean distance, and applying eq. 8 to obtain SS_{Total}, then eq. 3 to calculate BD_{Total}. How to choose an appropriate dissimilarity coefficient for a study is described in the next section.

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 or (dis)similarity coefficient should I use in my study? When assessing beta diversity using community composition variance, 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 coefficients

We describe four groups of properties and indicate the reason why we consider them relevant. The first two groups (i.e. from P1 to P7) contain the minimum requirements for assessing beta diversity. The remaining two groups (i.e. P8 to P14) are not necessarily required in all beta diversity assessments. Practitioners should determine whether the context of their analyses requires these properties or not. Similarity coefficients should be transformed into dissimilarities before assessing the following properties.

Property class 1: Basic necessary properties. —Properties P1 to P3 must be fulfilled by all resemblance coefficients used for beta diversity assessment. P1 and P2 are actually mathematical axioms that define a dissimilarity function. Thus, they are fulfilled by all coefficients considered in this paper and are therefore not shown in Table 1.

P1 – Minimum of zero and positiveness. A dissimilarity value should never be negative and it should be zero when comparing a site to itself. When comparing two different sites, it can be zero or greater than zero, depending on the species abundance values and how the dissimilarity is defined. For example, with some coefficients, *D* is zero when comparing two site vectors whose abundance values are proportional to each other; that is the case with the profile, chi-square, chord, and Hellinger distances. Dissimilarities that violate this property by taking negative values are called nonmetric, by opposition to the metric and semimetric coefficients described below.

P2 – Symmetry. Consider two community abundance vectors, $\mathbf{x}_1$ and $\mathbf{x}_2$, whose dissimilarity

is to be assessed. In symmetric indices, $D(\mathbf{x}_1,\mathbf{x}_2) = D(\mathbf{x}_2,\mathbf{x}_1)$. In the incidence-based counterparts of these coefficients (Table 1), the values *b* and *c* play identical (symmetric) roles. When studying beta diversity, there is no reason to make a distinction between the two sampling units that are compared using a coefficient. Therefore, dissimilarity coefficients must be symmetric. The property of being *double-zero symmetrical*, referred to in P3, is a different property.

P3 – Independence from double-zeros. Species that are absent in both sampling units produce double zeros in the data table. Double zeros in community composition data should not be interpreted, and should not affect dissimilarity coefficients, because a species may be absent from two sites either for the same reason (e.g. pH too high at both sites for that species) or for opposite reasons (e.g. pH too low at one site and too high at the other). The fact that some coefficients change their values in the presence of double-zeros in a comparison is referred to as the double-zero problem (Legendre & Legendre 2012, Section 7.2.2). Coefficients that produce smaller dissimilarity values when there are more double zeros in a comparison of sites are called *symmetrical* (or *double-zero symmetrical*) because they treat double zeros like any other pair of identical values. *Double-zero asymmetrical* coefficients, e.g. most of the coefficients discussed in the present paper, treat double absences in a different way than double presences; any number of double zeros do not change the values of these coefficients. Whereas P1 and P2 are mathematical axioms, P3 is, for ecological reasons, a necessary property for beta diversity studies.

Independence from double-zeros can easily be established for coefficients that are bounded by a maximum value (D_{max} in Table 2). For coefficients that do not have an upper bound, like the Euclidean and Manhattan distances, the fact that they are double-zero symmetrical (code 0 in column P3) is more difficult to establish. The demonstration is based on the binary form of the coefficient (Table 1, column 3), which is the one-complement of the

simple matching coefficient multiplied by the number of species p for the Manhattan distance, and its square root for the Euclidean distance. The simple matching coefficient is the archetype of double-zero symmetrical coefficients. The modified mean character difference is asymmetrical; its binary form is the Jaccard coefficient where double-zeros are explicitly excluded by division by $pp = a + b + c$.

Property class 2: Comparability between data sets. — The following properties are needed to appropriately compare beta diversity values calculated for different data tables, even if the sampling methods are the same.

P4 – Invariance to the number of sampling units (*n*) in the data table. Because it measures the average dispersion per sampling unit, variance is, by definition, independent of the number of sampling units in the data set. Similarly, if the variance estimate is computed from a dissimilarity coefficient through eqs 8 and 3, the number of sampling units in the data table should not influence the dissimilarity between pairs of sampling units. If that is not the case, the comparison of beta values between data tables of different sizes is not valid.

P5 – Invariance to the number of species in the data table. This property of incidence-based indices, referred to as *homogeneity* by several authors (e.g. Janson and Vegelius 1981, Koleff *et al.* 2003, Chao *et al.* 2006), allows the comparison of beta values computed from data tables containing different total species richness. We checked this property on the binary form of the coefficients, by multiplying a, b and c by a constant factor and checking whether the resulting index value was changed.

P6 – Invariance to the total abundance in the data table. This property concerns abundance-based formulas only. It allows the comparison of beta values between data tables (e.g. regions) with different productivities (abundance or biomass), or where biomass has been measured using different units (e.g. in g and mg). To see whether a given quantitative

coefficient is invariant to changes of measurement scale, we multiplied the abundance values by a constant factor and checked whether the resulting index was altered.

P7 – Existence of a fixed upper bound. The existence of an upper bound for a coefficient facilitates the interpretation and comparison of beta values because an upper bound in the dissimilarity coefficient leads to an upper bound in the beta diversity value. The maximum beta value for a region is obtained when all site pairs have the maximum dissimilarity D_{max} permitted by the chosen coefficient, and this happens when each site has no species in common with any other site. In practice, this also requires that $p \geq n$. One can apply eq. 8 to that situation to compute the maximum sum of squares:

$$SS_{max} = \frac{1}{n}\left(\frac{n(n-1)}{2} D^2_{max}\right) = \frac{n-1}{2} D^2_{max} \quad (11)$$

and then eq. 3 to obtain the maximum beta diversity value:

$$BD_{max} = \left(\frac{n-1}{2} D^2_{max}\right)\frac{1}{n-1} = \frac{1}{2} D^2_{max} \quad (12)$$

The upper bound varies among dissimilarity coefficients (Table 2). For coefficients with D_{max} = 1, BD_{max} = 0.5; if $D_{max} = \sqrt{2}$, BD_{max} = 1; and for the chi-square distance where $D_{max} = \sqrt{2y_{++}}$, $BD_{max} = y_{++}$ which is the sum of the species abundances in **Y**. Hence, for the coefficients that have a fixed maximum (see section *The dissimilarity measures*), we can compute a relative value of beta diversity, BD_{rel}, as follows:

$$BD_{rel} = BD_{Total} / BD_{max} \quad (13)$$

which is a value between 0 and 1. BD_{rel} is useful to compare beta values computed using different coefficients.

Property class 3: Sampling unit size, nestedness and undersampling. — This group of

properties is mostly related to sampling issues. The fulfilment of properties P8 and P9 facilitates (but does not ensure) the comparison of beta values obtained from sampling unit with different sizes. Indeed, both the number of species and the total abundance should be strongly affected by changes in the size of the sampling units. The remaining two properties deal with nestedness (P10) and correction for undersampling (P11) of the community composition. These properties are also related to sampling unit size, because small sampling units can lead to undersampling the targeted community richness, and differences in sampling unit sizes can produce nestedness in the community composition of sampling units, even when these come from the same community type.

P8 – Invariance to the number of species in each sampling unit. We checked this property on the incidence-based forms of the coefficients using the method described in Appendix S1 in Supporting Information.

P9 – Invariance to the total abundance in each sampling unit. Except when researchers only count and identify a fixed number of individuals (which is often the case in plankton or palaeoecological studies), sampling units in the data table are likely to have different total abundances. Some abundance-based dissimilarity indices are only sensitive to relative abundances per site whereas others reflect differences in site total counts. This property was called "density invariance" by Jost *et al.* (2011). It is not the same as property P6 above. One can check property P9 by determining whether a coefficient is altered when the abundances are multiplied by a constant factor that is different for each sampling unit.

P10 – Zero-value for nested species composition. If all species in a sampling unit are also found in another sampling unit (that is, if either $b = 0$ or $c = 0$ but not both), some dissimilarity coefficients produce a zero value (i.e. they are nestedness-independent) whereas most others do not. When both b and c are 0, all coefficients return a dissimilarity of 0. Some

authors consider that separating nestedness from species replacement is important for beta diversity assessment (Baselga 2010).

P11 – Coefficients with corrections for undersampling. With higher sampling effort, i.e. larger sampling units, rare species, and in particular those that are not found at the two sites under comparison, are more likely to be observed (Chao *et al.* 2006, Cardoso *et al.* 2009). For that reason, dissimilarity coefficients generally underestimate the dissimilarities among sites, the bias decreasing when sampling effort increases. For some binary similarity coefficients, Chao *et al.* (2006) and Jost *et al.* (2011) suggested abundance-based counterparts that incorporate corrections for undersampling bias.

Property class 4: Ordination-related properties. — The remaining properties are not related to the ecological interpretation of a coefficient or the comparability of beta diversity values. They are, however, useful for ordination of community composition data.

P12 metric and P13 Euclidean properties of D or $\mathbf{D}^{(0.5)}$. A dissimilarity matrix **D** is *metric* if it has the following metric properties: positiveness (P1), symmetry (P2), and triangle inequality, i.e. for any triplet of distance values, $D(\mathbf{x}_1,\mathbf{x}_2)+D(\mathbf{x}_2,\mathbf{x}_3)\geq D(\mathbf{x}_1,\mathbf{x}_3)$. Metric dissimilarities are also called distances. **D** is *Euclidean* if it can be embedded in an Euclidean space of real axes such that the Euclidean distances among points are equal to the dissimilarity values in **D**. When this property is satisfied, ordination by principal coordinate analysis of **D** does not generate negative eigenvalues. Gower and Legendre (1986) checked the metric and Euclidean properties of several binary and quantitative coefficients. The original dissimilarity coefficient may have the metric property (P12). Coefficients that have properties P1 and P2 but may violate the triangle inequality property are not metric; they are called *semimetric*. For some of these, metricity can be obtained by taking the square root of the dissimilarity values; the resulting matrix, which contains values $\left[D_{ij}^{0.5}\right]$, is noted $\mathbf{D}^{(0.5)}$.

Likewise for the Euclidean property (P13). Legendre & Legendre (2012, Tables 7.2 and 7.3) describe the metric and Euclidean properties of 43 commonly-used similarity and dissimilarity coefficients, including several of the coefficients listed in Table 1.

P14 – Emulated by transformation of the raw frequency data. Legendre & Gallagher (2001) described how some distance coefficients can be obtained by computing the Euclidean distance (eq. 7) after transforming the raw data values in some appropriate way. Another distance coefficient can be obtained by applying the Manhattan distance to transformed data; see next section. The transformations are described in Appendix S2. These transformations are important because they allow one to easily compute a dissimilarity coefficient and, at the same time, identify the beta diversity contributions of individual species through eqs 4a and 4b, and of sites through eqs 5a and 5b. One can also use the transformed data directly in linear modelling of community composition data, e.g. by simple (PCA) or canonical (RDA) ordination, *K*-means partitioning, or multivariate regression tree analysis (MRT), because these methods implicitly preserve the Euclidean distance among sites. They are useful to describe and explain the observed patterns of beta diversity (Legendre & Legendre 2012, Section 7.7).

The dissimilarity coefficients

A selection of 17 quantitative dissimilarity coefficients commonly used for beta diversity assessment was considered in the comparison (Table 1). They represent a broad hand among the available coefficients. Equations are shown for community composition abundance and also for presence-absence (i.e. incidence) data. The properties (P3 to P14) of these coefficients are listed in Table 2, as well as their D_{max} values when they exist.

The first 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 P3) 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 S2). The Manhattan distance, which is also inappropriate *per se*, is included in the comparison because, like the Euclidean distance, it may become appropriate after data transformation.

The other coefficients included in the comparative study are double-zero asymmetrical (property P3); they have been recommended and used for community composition assessment or beta diversity studies. Two groups of coefficients deserve additional comments.

Five of the distances 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 S2); computing the Euclidean distance (eq. 7) on the transformed data produces the targeted distance. In the same way, computing the Manhattan distance on data transformed into species profiles produces twice Whittaker's index of association; for that reason, Whittaker's index was dubbed "relativized Manhattan" by Faith *et al.* (1987). Actually, the four transformations described in Appendix S2 could be used before calculation of the Manhattan distance.

A consequence of this observation is that SS_{Total} corresponding to the species profile, Hellinger, chord, and chi-square distances can be obtained by computing the transformation in Appendix S2, then applying eqs 1 and 2 to the transformed data. This is simpler than computing the distance matrix, then using eq. 8 to obtain SS_{Total}. Furthermore, it allows the computation of SCBD statistics (eq. 4b), which cannot be obtained from a distance matrix.

The last four dissimilarities in Table 1, proposed by Chao *et al.* (2006), implement corrections for undersampling bias (P11). These coefficients are not Euclidean 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 (P13) when transformed to $D = \sqrt{1 - \text{similarity}}$ (Legendre & Legendre 2012, Table 7.2).

When applied to presence-absence data, several quantitative dissimilarity functions in Table 1 produce either the one-complement of the Jaccard similarity index, $(b+c)/(a+b+c)$, or the one-complement of the Sørensen index, $(b+c)/(2a+b+c)$. 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 P3 to P14 as columns. Most properties were coded as presence-absence (0-1), except for coefficients P12 and P13 which were coded on a semiquantitative 0-1-2 scale (0 = property absent, 1 = present for $\mathbf{D}^{(0.5)}$, 2 = present for $\mathbf{D}^{(0.5)}$ and **D**). NA values in Table 2 were treated as zeros since the coefficient did not have the property in question. 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 P3 to P14, which formed the variables of the matrix subjected to PCA, are shown as red arrows. One can identify five types of coefficients in the ordination diagram:

• Type I contains the Euclidean and Manhattan distances as well as the mean character difference. The first two coefficients lack property P3 (asymmetric treatment of double-zeros)

which makes them unsuitable for beta diversity assessment. All three lack P6 (invariance with respect to y_{++}) and P7 (fixed upper bound); these coefficients would not allow comparison of beta diversity estimates among data sets.

Coefficients in types II to V provide asymmetrical treatment of double-zeros (P3). They also all have properties P6 and P7, which are required for comparability of beta estimates among data sets. They are thus all appropriate for beta diversity assessment.

• Type II contains the profile, Hellinger, chord, chi-square and Whittaker distances. They share properties P12 (metric), P13 (Euclidean) and P14 (emulated by transformation of the raw frequency data). **D** matrices computed using these coefficients ($\mathbf{D}^{(0.5)}$ in the Whittaker case) are fully suitable for ordination by principal coordinate analysis (PCoA), which will not produce negative eigenvalues and complex axes. Furthermore, species frequency (or frequency-like, such as biomass) data transformed using the profile, Hellinger, chord, or chi-square transformations can be used directly in principal component analysis (PCA) and in canonical ordination by RDA; this is not the case for type III-V coefficients.

• Type III contains the divergence, Canberra, percentage difference (*alias* Bray-Curtis), and Wishart dissimilarities. The coefficient of divergence, which is Euclidean, can be used directly in PCoA ordination. For the other three coefficients, 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).

• Type IV contains the abundance-based quantitative forms of the Jaccard, Sørensen, Ochiai and Simpson indices. They share properties P9 (invariance to total abundance in individual sampling unit) and P11 (correction for undersampling), but not properties P12, P13 and P14, which are desirable for ordination.

• The Kulczynski coefficient forms a type of its own (type V). It is suitable for beta diversity assessment, but not for ordination, and it does not correct for undersampling. This coefficient does not offer any particular advantage not available in other coefficients.

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 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 S2).

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 and the relative contributions ($LCBD_i$, eq. 5b) were in the range [0.0191, 0.0637]; LCBD indices, which indicate the uniqueness of the fish community at each site, are plotted on a map of the river in Fig. 3a. Comparison with species richness (Fig. 3b) showed that LCBD was negatively related 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.

Environmental variables were also available for each site. They describe distance from the source, altitude, 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. Only two environmental variables were retained by backward elimination: slope of the riverbed and BOD. The regression model had an adjusted R^2 of 0.58; both variables had positive coefficients in the model, indicating that sites contributing highly to 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 ordination 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 of 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. the BD_{Total} statistic, into explained and residual components. By contrast, in regression analysis of the LCBD vector, the question is why some sites have higher degrees of uniqueness in species composition than others.

Five species contributed to beta diversity more than the mean of the 27 species: the common roach (Cyprinidae) with the lowest value of SCBD above the mean, the stone loach (Balitoridae), the common bleak (Cyprinidae), the Eurasian minnow (Cyprinidae), and the brown trout (Salmonidae) with the highest SCBD value. The chord-transformed abundances of these species varied the most among the sites. The brown trout, Eurasian minnow and stone loach are found in the high LCBD sites upriver, whereas the common roach and common bleak are abundant in the eutrophic sites in the middle course of the river that also have high LCBD values.

Calculation of LCBD was repeated using the 17 coefficients in Table 1 using the software in Appendix S3. A cluster analysis was carried out from the Spearman correlation matrix among the 17 LCBD vectors. The results (Fig. 4) indicate that the LCBD vectors

computed using the 14 coefficients pertaining to types II to V form a large cluster separated from the first three coefficients. LCBDs were quite homogeneous in that large group: the mean Spearman correlation among the 14 LCBD vectors was 0.869. Kendall concordance analysis with *a posteriori* tests (Legendre 2005) showed that the contributions of all 14 vectors to the concordance of the large 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.) We conclude that, for this example, LCDB indices computing using all dissimilarities that were found suitable for beta diversity assessment were highly concordant.

DISCUSSION

Related approaches to beta diversity assessment

In this paper, we used the total variance of **Y** as an estimate of beta diversity (DB_{Total}) for a region of interest (eq. 3). Alternative equations have been proposed by Whittaker (1972), Ricotta & Marignani (2007) and Anderson *et al.* (2006). We will now show that these proposals are special cases of eq. 3. In section "Equivalent ways of computing Var(**Y**)", we saw that SS(**Y**) can be computed as the sum of the squared dissimilarities divided by *n* (eq. 8). This is appropriate for the Euclidean distance and also for dissimilarities that have the property of being Euclidean (P13). Several dissimilarity indices, coded 1 for property P13 in Table 2, are Euclidean only when taking their square roots; the transformed distances form matrix $\mathbf{D}^{(0.5)} = \left[D_{ij}^{0.5}\right]$. That group includes Whittaker's index, the Canberra metric, the widely-used percentage difference (*alias* Bray-Curtis) and Wishart's coefficient. Many of the incidence-based (i.e. binary) coefficients used in community ecology are also in that situation, including the widely-used Jaccard, Sørensen and Ochiai coefficients (Legendre & Legendre

2012, Table 7.2). The method of calculation of beta diversity proposed in other papers is equivalent to DB_{Total} of the present paper if $\mathbf{D}^{(0.5)}$ is used for the calculations:

(a) Whittaker (1972, p. 233) stated that "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". The mean is obtained by summing the dissimilarities and dividing by the number of dissimilarities in the half-matrix, $n(n-1)/2$. This is equivalent to computing eqs 8 and 3 on the square-rooted dissimilarities (matrix $\mathbf{D}^{(0.5)}$) and multiplying by 2. Hence, Whittaker's formula only differs by a factor 2 from DB_{Total} computed from $\mathbf{D}^{(0.5)}$.

(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 plots 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. It differs from DB_{Total}, which is *the sum of the squared dissimilarities* from the sampling units to the group centroid (eq. 10a) divided by $(n-1)$ (eq. 3). Because it is computed from any dissimilarity matrix (eqs 9 and 10a; the square root of the values in vector $[SS_i]$ provide the dissimilarities of the sampling units to the centroid), the Anderson *et al.* (2006) statistic can be obtained from $\mathbf{D}$ as well as $\mathbf{D}^{(0.5)}$.

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 **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. The beta statistic of Anderson *et al.* (2006), which is closely related to DB_{Total}, can be computed from **D** or $\mathbf{D}^{(0.5)}$. (b) If the dissimilarity measure can be obtained by applying a transformation to the original data (Appendix S2) 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).

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 Var(**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). This can be done whether the calculation of SS_{Total} has been done from the raw or transformed abundance data. The centred SCBD values, which are signed, indicate the species that vary more [or less] than the mean across the sites.

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 represents site uniqueness; hence, large LCBD values indicate sites that have strongly different species compositions. In conservation biology, LCBD could be used as an indicator of the site conservation value. LCBD may be inversely correlated with species richness, as in our example, but in other ecosystems large LCBDs 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 small 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. *Spatial scales*. — SS_{Total} can be partitioned as a function of different 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).

6. *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.

7. *Multivariate variogram and ordination*. — SS_{Total} can be partitioned into spatial scales by multivariate variogram analysis (Wagner 2003). The species-environment relation, which represents a portion of SS_{Total}, can also 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

Analyzing 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 17 selected dissimilarity coefficients based on 14 ecological, statistical and mathematical properties led to a model where the coefficients were divided into five main types. Four 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. 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

Our thanks to Daniel Borcard who provided comments on the manuscript before submission. 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 b 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.

Bloom, S.A. (1981). Similarity indices in community studies: potential pitfalls. *Mar. Ecol. Progr. Ser.*, 5, 125–128.

6.

Borcard, D., Gillet, F. & Legendre, P. (2011). *Numerical ecology with R.* Use R! series, Springer Science, New York.

7.

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

8.

Bray, R.J. & Curtis, J.T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr.*, 27, 325–349.

9.

Cardoso, P., Borges, P.A.V. & Veech, J.A. (2009). Testing the performance of beta diversity measures based on incidence data: the robustness to undersampling. *Divers. Distrib.*, 15, 1081–1090.

10.

Chao, A., Chazdon, R.L., Colwell, R.K. & Shen, T.J. (2006). Abundance-based similarity indices and their estimation when there are unseen species in samples. *Biometrics* 62, 361–371.

11.

Chao, A., Chiu, C.-H. & Hsieh, T.C. (2012). Proposing a resolution to debates on diversity partitioning. *Ecology*, 93, 2037–2051.

12.

Clark, P.J. (1952). An extension of the coefficient of divergence for use with multiple characters. *Copeia*, 1952, 61–64.

13.

Clarke, K.R., Somerfield, P.J. & Chapman, M.G. (2006). On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray–Curtis coefficient for denuded assemblages. *J. Exp. Mar. Biol. Ecol.*, 330, 55–80.

14.

Czekanowski, J. (1909). Zur Differentialdiagnose der Neandertalgruppe. *Korrespondenz-Blatt deutsch. Ges. Anthropol. Ethnol. Urgesch.*, 40, 44–47.

15.

De Cáceres, M., Legendre, P., Valencia, R., Cao, M., Chang, L.-W., Chuyong, G. *et al.* (2012). The variation of tree beta diversity across a global network of forest plots. *Global Ecology and Biogeography*. DOI: 10.1111/j.1466-8238.2012.00770.x

16.

Dray, S., Pélissier, R., Couteron, P., Fortin, M.-J., Legendre, P., Peres-Neto, P.R. *et al.* (2012). Community ecology in the age of multivariate multiscale spatial analysis. *Ecol. Monogr.*, 82, 257–275.

17.

Ellison, A.M. (2010). Partitioning diversity. *Ecology*, 91, 1962–1963.

18.

Faith, D.P., Minchin, P.R. & Belbin, L. (1987). Compositional dissimilarity as a robust measure of ecological distance. *Vegetatio*, 69, 57–68.

19.

Gower, J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53, 325–338.

20. Gower, J.C. & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *J. Classif.*, 3, 5–48.

21. Hajdu, L.J. (1981). Graphical comparison of resemblance measures in phytosociology. *Vegetatio*, 48, 47–59.

22. Hubálek, Z. (1982). Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation. *Biol. Rev.*, 57, 669–689.

23. Hurlbert, S.H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecol. Monogr.*, 54, 187–211.

24. Jaccard, P. (1900). Contribution au problème de l'immigration post-glaciaire de la flore alpine. *Bull Soc. Vaudoise Sci. Nat.*, 36, 87–130.

25. Janson, S. & Vegelius, J. (1981). Measures of ecological association. *Oecologia*, 49, 371–376.

26. Janssen, J.G.M. (1975). A simple clustering procedure for preliminary classification of very large sets of phytosociological relevés. *Vegetatio*, 30, 67–71.

27. Jost, L. (2007). Partitioning diversity into independent alpha and beta components. *Ecology*, 88, 2427–2439.

28.

734 Jost, L., Chao, A. & Chazdon, R.L. (2011). Compositional similarity and beta diversity. In:
735 *Biological diversity: frontiers in measurement and assessment* [eds Magurran, A. &
736 McGill, B.]. Oxford University Press, Oxford, England, pp. 66–84.
737 29.
738 Koleff, P., Gaston, K.J. & Lennon, J.J. (2003). Measuring beta diversity for presence-absence
739 data. *J. Anim. Ecol.*, 72, 367–382.
740 30.
741 Kraft, N.J.B., Comita, L.S., Chase, J.M., Sanders, N.J., Swenson, N.G., Crist, T.O. *et al.*
742 (2011). Disentangling the drivers of diversity along latitudinal and elevational
743 gradients. *Science*, 333, 1755–1758.
744 31.
745 Kulczynski, S. (1928). Die Pflanzenassoziationen der Pieninen. *Bull. Int. Acad. Pol. Sci. Lett.*
746 *Cl. Sci. Math. Nat. Ser. B*, Suppl. II (1927), 57–203.
747 32.
748 Lance, G.N. & Willams, W.T. (1967). Mixed-data classificatory programs. I. Agglomerative
749 systems. *Aust. Comput. J.*, 1, 15–20.
750 33.
751 Lebart, L. & Fénelon, J.P. (1971). *Statistique et informatique appliquées.* Dunod, Paris, France.
752 34.
753 Legendre, P. (2005). Species associations: the Kendall coefficient of concordance revisited. *J.*
754 *Agr. Biol. Envir. S.*, 10, 226–245.
755 35.
756 Legendre, P. & Anderson, M.J. (1999). Distance-based redundancy analysis: testing
757 multispecies responses in multifactorial ecological experiments. *Ecol. Monogr.*, 69, 1–
758 24.

36.

Legendre, P., Borcard, D. & Peres-Neto, P.R. (2005). Analyzing beta diversity: partitioning the spatial variation of community composition data. *Ecol. Monogr.*, 75, 435–450.

37.

Legendre, P. & Fortin, M.-J. (2010). Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Mol. Ecol. Resour.*, 10, 831–844.

38.

Legendre, P. & Gallagher, E.D. (2001). Ecologically meaningful transformations for ordination of species data. *Oecologia*, 129, 271–280.

39.

Legendre, P. & Legendre, L. (2012). *Numerical ecology*. 3rd English edition. Elsevier Science BV, Amsterdam.

40.

Odum, E.P. (1950). Bird populations of the Highlands (North Carolina) Plateau in relation to plant succession and avian invasion. *Ecology*, 31, 587–605.

41.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B. *et al.* (2012). *vegan: Community ecology package. R package version 2.0-3.* Available at: http://cran.r-project.org/web/packages/vegan/.

42.

Orlóci, L. (1967). An agglomerative method for classification of plant communities. *J. Ecol.*, 55, 193–206.

43.

Orlóci, L. (1978). *Multivariate analysis in vegetation research.* 2nd edition. Dr. W. Junk B.

V., The Hague, The Netherlands.

44.

Pelissier, R., Couteron, P., Dray, S. & Sabatier, D. (2003). Consistency between ordination techniques and diversity measurements: two strategies for species occurrence data. *Ecology*, 84, 242–251.

45.

Peres-Neto, P.R., Legendre, P., Dray, S. & Borcard, D. (2006). Variation partitioning of species data matrices: estimation and comparison of fractions. *Ecology*, 87, 2614–2625.

46.

Rao, C.R. (1982). Diversity and dissimilarity coefficients: a unified approach. *Theor. Popul. Biol.*, 21, 24–43.

47.

Rao, C.R. (1995). A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. *Qüestiió (Quaderns d'Estadistica i Investivació Operativa)*, 19, 23–63.

48.

Ricotta, C. & Marignani, M. (2007). Computing B-diversity with Rao's quadratic entropy: a change of perspective. *Divers. Distrib.*, 13, 237–241.

49.

Simpson, G.G. (1943). Mammals and the nature of continents. *Am. J. Sci.*, 241, 1–31.

50.

Stephenson, W., Williams, W.T. & Cook, S.D. (1972). Computer analyses of Petersen's original data on bottom communities. *Ecol. Monogr.*, 42, 387–415.

51.

Vellend, M. (2001). Do commonly used indices of beta-diversity measure species turnover? *J. Veg. Sci.*, 12, 545–552.

52.

Verneaux, J. (1973). Cours d'eau de Franche-Comté (Massif du Jura). Recherches écologiques sur le réseau hydrographique du Doubs – Essai de biotypologie. *Annales Scientifiques de l'Université de Franche-Comté, Biologie Animale*, 3, 1–260.

53.

Wagner, H.H. (2003). Spatial covariance in plant communities: integrating ordination, variogram modeling, and variance testing. *Ecology*, 84, 1045–1057.

54.

Wagner, H.H. (2004). Direct multi-scale ordination with canonical correspondence analysis. *Ecology*, 85, 342–351.

55.

Whittaker, R.H. (1952). A study of summer foliage insect communities in the Great Smoky Mountains. *Ecol. Monogr.*, 22, 1–44.

56.

Whittaker, R.H. (1960). Vegetation of the Siskiyou mountains, Oregon and California. *Ecol. Monogr.*, 30, 279–338.

57.

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

58.

Wilson, M.V. & Shmida, A. (1984). Measuring beta diversity with presence-absence data. *J. Ecol.*, 72, 1055–1064.

59.

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

Andrews, St. Andrews, Fife, Scotland.

SUPPORTING INFORMATION

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

Appendix S1 Details about property P8: Invariance to the number of species in each sampling unit.

Appendix S2 Community composition data transformations.

Appendix S3 The R function beta.div() computes estimates of total beta diversity as the total variance in a community data matrix **Y**, as well as the derived SCBD and LCBD statistics, for 17 dissimilarity coefficients or the raw data table.

Table 1 Dissimilarity coefficients compared in this paper.

Dissimilarity	*Abundance-based*	*Incidence-based*	*References*	*Coefficient number in L&L* [1]
Euclidean distance	$\sqrt{\sum_{j=1}^{p}\left[y_{1j}-y_{2j}\right]^2}$	$\sqrt{p\left(\frac{b+c}{a+b+c+d}\right)}=\sqrt{b+c}$		D_1
Manhattan distance	$\sum_{j=1}^{p}\left\|y_{1j}-y_{2j}\right\|$	$p\left(\frac{b+c}{a+b+c+d}\right)=b+c$		D_7
Modified mean character difference	$\frac{1}{pp}\sum_{j=1}^{p}\left\|y_{1j}-y_{2j}\right\|$	$\frac{b+c}{a+b+c}$	Legendre & Legendre (2012)	D_{19}
Species profile distance	$\sqrt{\sum_{j=1}^{p}\left[\frac{y_{1j}}{y_{1+}}-\frac{y_{2j}}{y_{2+}}\right]^2}$	$\frac{\sqrt{a(b-c)^2+b(a+c)^2+c(a+b)^2}}{(a+b)(a+c)}$	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}
Whittaker's index of association	$\frac{1}{2}\sum_{j=1}^{p}\left\|\frac{y_{1j}}{y_{1+}}-\frac{y_{2j}}{y_{2+}}\right\|$	$\frac{a\left\|b-c\right\|+b(a+c)+c(a+b)}{2(a+b)(a+c)}$	Whittaker (1952)	D_9

Coefficient of divergence	$\sqrt{\frac{1}{pp}\sum_{j=1}^{p}\left(\frac{y_{1j}-y_{2j}}{y_{1+}+y_{2+}}\right)^2}$	$\sqrt{\frac{b+c}{a+b+c}}$	Clark (1952)	D_{11}
Canberra metric [3]	$\frac{1}{pp}\sum_{j=1}^{p}\frac{\lvert y_{1j}-y_{2j}\rvert}{(y_{1+}+y_{2+})}$	$\frac{b+c}{a+b+c}$	Lance & Willams (1967), Stephenson *et al.* (1972) for $1/pp$	D_{10}
Percentage difference (*alias* Bray-Curtis dissimilarity [4])	$\frac{\sum_{j=1}^{p}\lvert y_{1j}-y_{2j}\rvert}{y_{1+}+y_{2+}}$	$\frac{b+c}{2a+b+c}$	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	$\left(1-\frac{UV}{U+V-UV}\right)$	$\frac{b+c}{a+b+c}$	Chao *et al.* (2006)	
Abundance-based Sørensen	$\left(1-\frac{2UV}{U+V}\right)$	$\frac{b+c}{2a+b+c}$	Chao *et al.* (2006)	
Abundance-based Ochiai	$\left(1-\sqrt{UV}\right)$	$\left(1-\frac{a}{\sqrt{(a+b)(a+c)}}\right)$	Chao *et al.* (2006)	
Abundance-based Simpson	$\left(1-\frac{UV}{UV+\min(U-UV,V-UV)}\right)$	$\frac{\min(b,c)}{a+\min(b,c)}$	Simpson (1943), Chao *et al.* (2006)	

[1] L&L = Legendre & Legendre (2012, Chapter 7).

[4] NA: No binary form for this coefficient.

[3] Division by *pp* in the Canberra metric was introduced by Stephenson *et al.* (1972) and adopted by Oksanen *et al.* (2012).

[4] Coefficient first described by Steinhaus in 1940's, then by Odum (1950) as the *percentage difference*. The Bray & Curtis (1957) paper was about a new ordination method; the authors did not describe this coefficient in their paper, where they only used a simplified form of the coefficient. It is thus incorrect to attribute this coefficient to these authors. More details about this in Bray & Curtis (1957) and Legendre & Legendre (2012, p. 311).

Table 2 Properties P3 to P14 of the coefficients listed in Table 1, described in the text. 1 indicates that a coefficient has the property, 0 that it does not. For P12 (metric) and P13 (Euclidean), 2 indicates that both **D** and $\mathbf{D}^{(0.5)}$ have the property, 1 that only $\mathbf{D}^{(0.5)} = [D_{ij}^{0.5}]$ has it, and 0 that neither **D** nor $\mathbf{D}^{(0.5)}$ have it. NA: there is no binary form for the chi-square distance; hence, P5, P8 and P10 could not be assessed. Last column: maximum possible dissimilarity value (D_{max}).

Dissimilarity	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	D_{max}
Euclidean distance	0	1	0	0	0	0	0	0	0	2	2	0	—
Manhattan distance	0	1	0	0	0	0	0	0	0	2	1	0	—
Modified mean character difference	1	1	1	0	0	0	0	0	0	0	0	0	—
Species profile distance	1	1	0	1	1	0	1	0	0	2	2	1	$\sqrt{2}$
Hellinger distance	1	1	1	1	1	1	1	0	0	2	2	1	$\sqrt{2}$
Chord distance	1	1	1	1	1	1	1	0	0	2	2	1	$\sqrt{2}$
Chi-square distance	1	0	NA	1	1	NA	0	NA	0	2	2	1	$\sqrt{2y_{++}}$
Whittaker's index of association	1	1	1	1	1	0	1	0	0	2	1	1	1

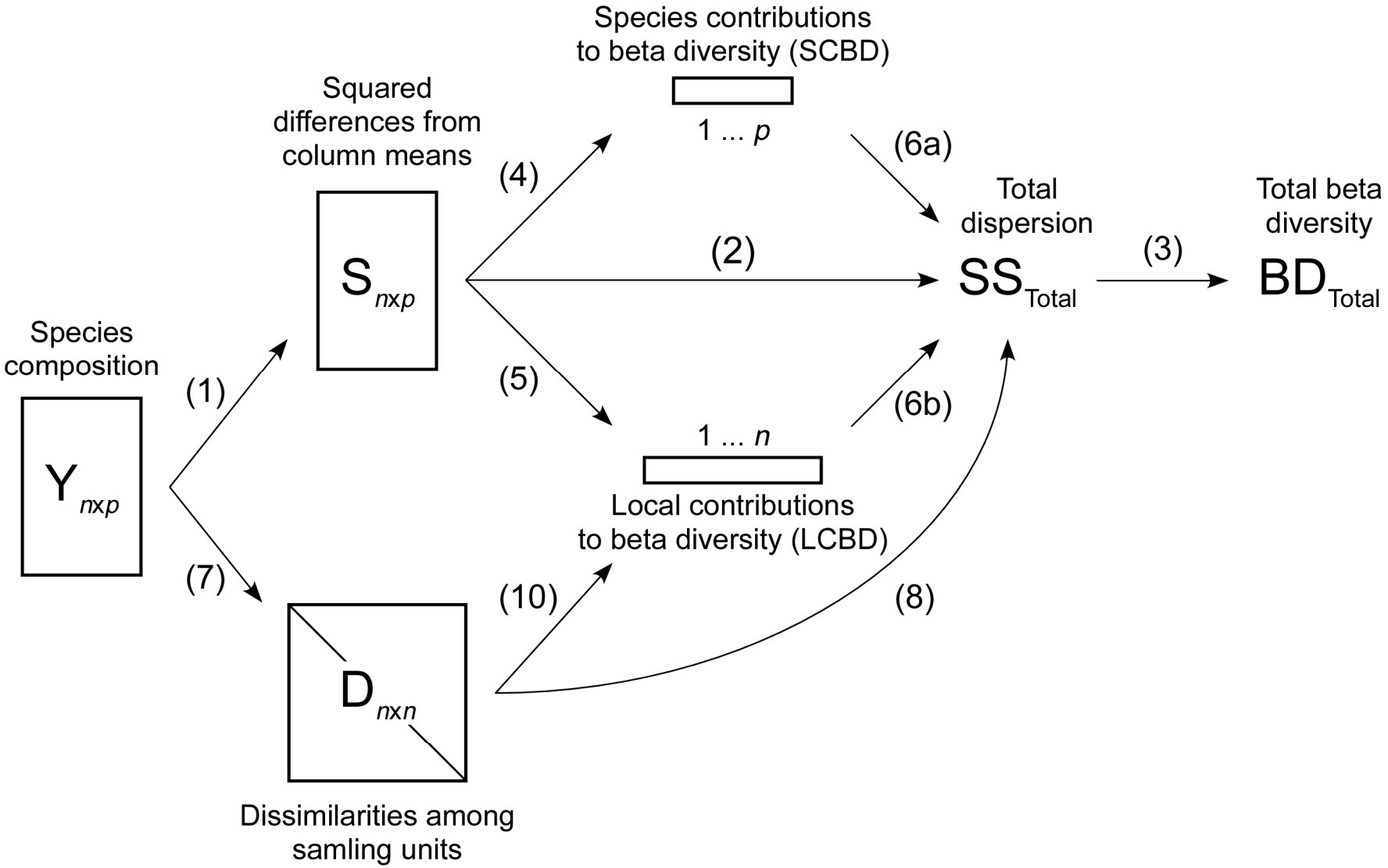
Coefficient of divergence	1	1	1	1	1	0	0	0	0	2	2	0	1
Canberra metric	1	1	1	1	1	0	0	0	0	2	1	0	1
Percentage difference (*alias* Bray-Curtis)	1	1	1	1	1	0	0	0	0	1	1	0	1
Wishart coefficient = (1 – similarity ratio)	1	1	1	1	1	0	0	0	0	1	1	0	1
D = (1–Kulczynski coefficient)	1	1	1	1	1	1	0	0	0	0	0	0	1
Abundance-based Jaccard	1	1	1	1	1	0	1	0	1	0	0	0	1
Abundance-based Sørensen	1	1	1	1	1	0	1	0	1	0	0	0	1
Abundance-based Ochiai	1	1	1	1	1	1	1	0	1	0	0	0	1
Abundance-based Simpson	1	1	1	1	1	0	1	1	1	0	0	0	1

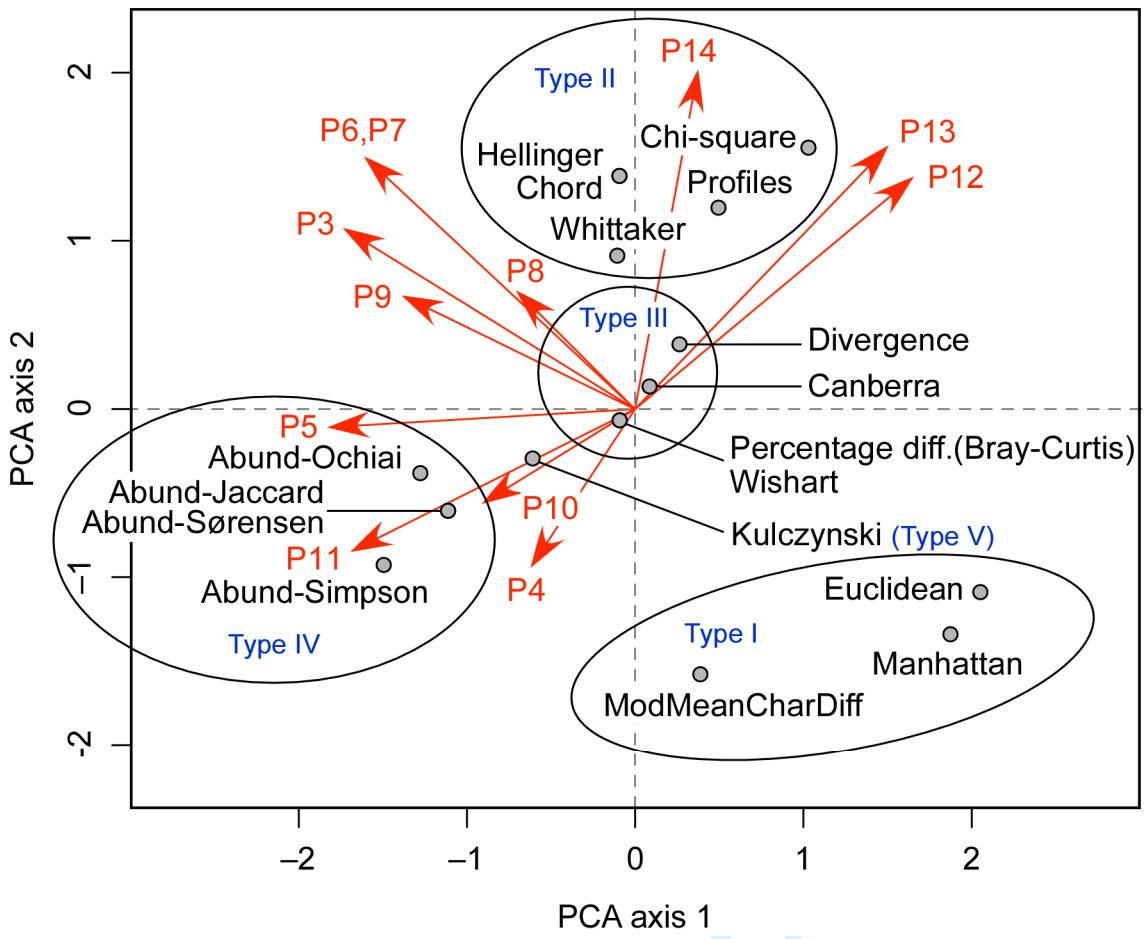
Figure 1 Schematic diagram representing the different ways of computing beta diversity as the total variance in **Y** as well as the contributions of individual species and sampling units, starting from the species composition data table **Y**. Numbers in parentheses refer to the equations in the text.

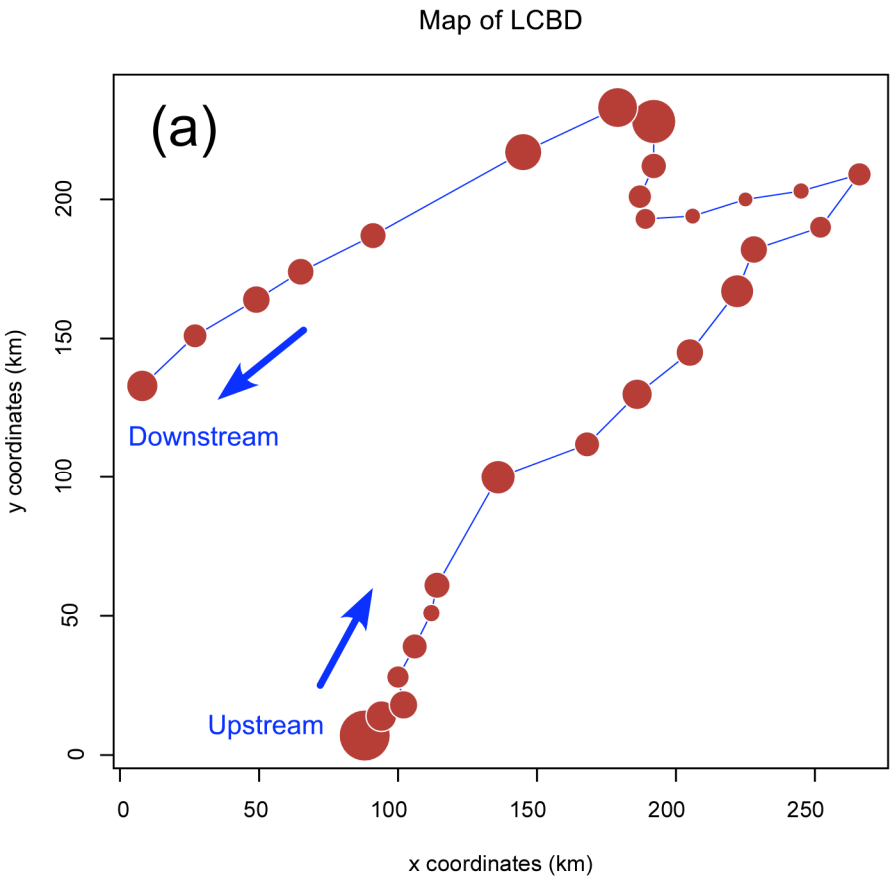
Figure 2 Principal component biplot showing the dissimilarity coefficients (gray points; see Table 1 for the full coefficient names) and properties P3 to P14 (red arrows). Examination of the grouping of coefficients (points) in the biplot and their properties (arrows) indicate five main types of coefficients. PCA axis 1 accounts for 34.8% of the multivariate variation and axis 2 for 25.5%.

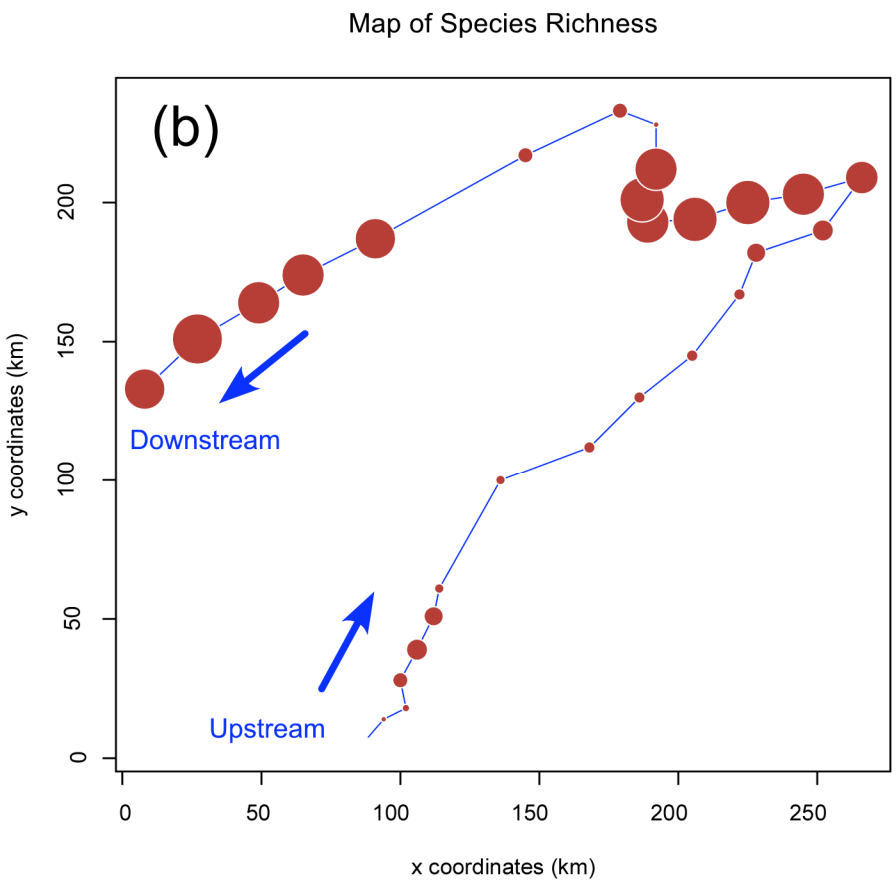
Figure 3 Maps of Doubs River (blue line) showing (a) the local contributions to beta diversity (LCBD) of the fish assemblage data and (b) the species richness at the 29 study sites. Size of the circles is proportional to the LCBD or richness values. The arrows indicate flow direction.

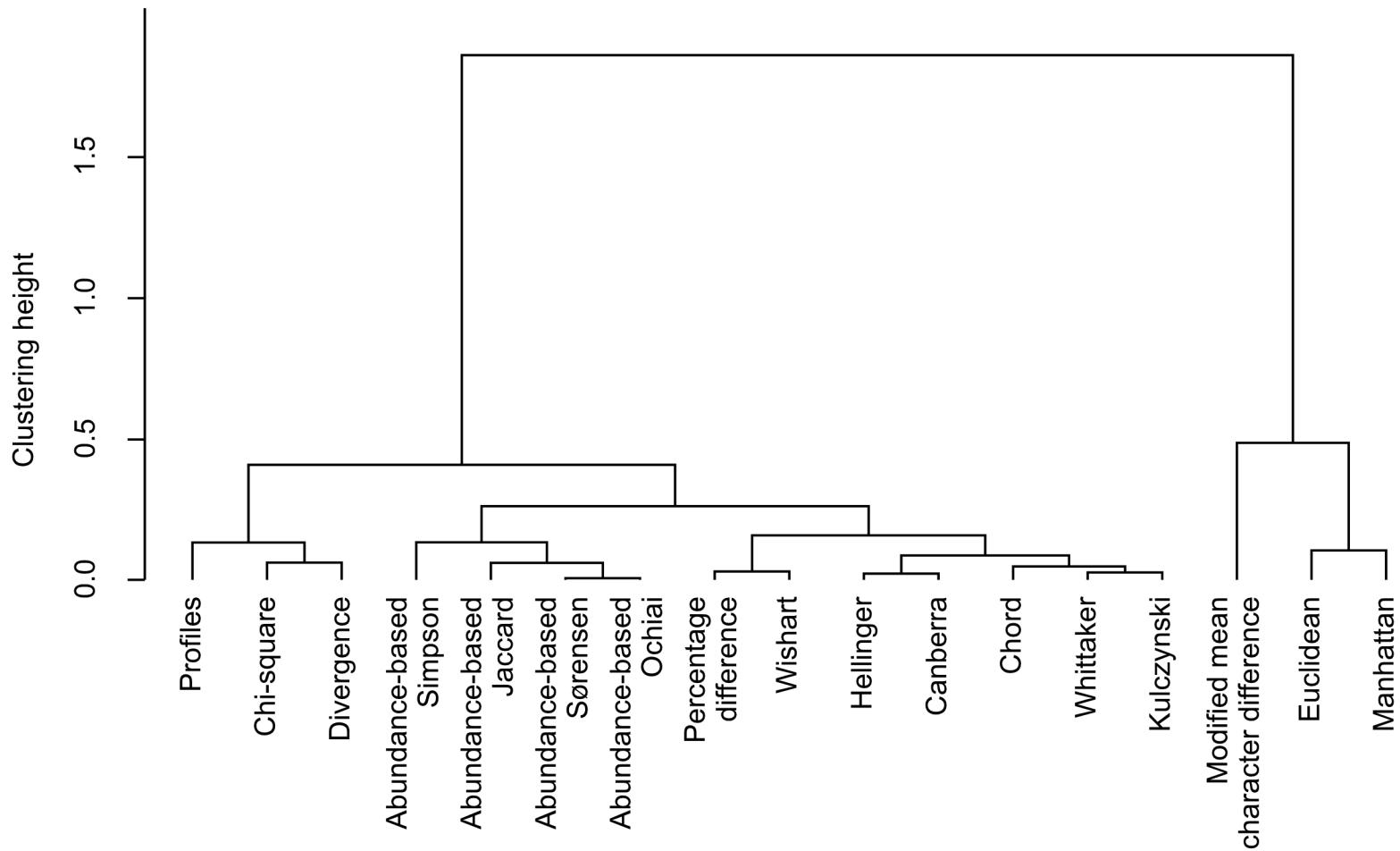
Figure 4 Cluster analysis dendrogram (Ward's method) of the local contributions to beta diversity (LCBD) vectors obtained after running the Doubs fish data through the 17 dissimilarity coefficients listed in Tables 1 and 2.


Species contributions
to beta diversity (SCBD)
1 ... p
Squared
differences from
column means
(4)
(6a)
Total
dispersion
Total beta
diversity
(2)
(3)
S nxp
SS Total
BD Total
Species
composition
(1)
(5)
(6b)
1 ... n
Y nxp
Local contributions
to beta diversity (LCBD)
(7)
(10)
(8)
D nxn
Dissimilarities among
samling units


Type II
P14
P6,P7
Hellinger
Chord
Chi-square
Profiles
Whittaker
P13
P12
P3
P8
P9
Type III
Divergence
Canberra
P5
Abund-Ochiai
Abund-Jaccard
Abund-Sørensen
P11
Abund-Simpson
Type IV
P10
P4
Percentage diff.(Bray-Curtis)
Wishart
Kulczynski (Type V)
Euclidean
Type I
Manhattan
ModMeanCharDiff
PCA axis 2
PCA axis 1
2
1
0
−1
−2
−2
−1
0
1
2


Map of LCBD
(a)
Downstream
Upstream
y coordinates (km)
x coordinates (km)
0
50
100
150
200
250


Map of Species Richness
(b)
Downstream
Upstream
y coordinates (km)
x coordinates (km)
0
50
100
150
200
250


Clustering height
0.0
0.5
1.0
1.5
Profiles
Chi-square
Divergence
Abundance-based Simpson
Abundance-based Jaccard
Abundance-based Sørensen
Abundance-based Ochiai
Percentage difference
Wishart
Hellinger
Canberra
Chord
Whittaker
Kulczynski
Modified mean character difference
Euclidean
Manhattan