Research

Oikos 118: 391–402, 2009
doi: 10.1111/j.1600-0706.2009.16668.x,

Subject Editor: Owen Petchey. Accepted 30 September 2008

On the challenge of treating various types of variables: application for improving the measurement of functional diversity

Sandrine Pavoine, Jeanne Vallet, Anne-Béatrice Dufour, Sophie Gachet and Hervé Daniel

S. Pavoine (sandrine.pavoine@zoo.ox.ac.uk), Muséum National d'Histoire Naturelle, Dépt Ecologie et Gestion de la Biodiversité, UMR 5173 MNHN-CNRS-P6, 55 rue Buffon, FR–75005 Paris, France. Present address: Dept of Zoology, Univ. of Oxford, South Parks Road, Oxford OX1 3PS, UK. – J. Vallet and H. Daniel, Agrocampus Ouest, Centre d'Angers, Inst. National d'Horticulture et de Paysage, UP Paysage, 2 rue Le Nôtre, FR–49045 Angers Cedex 01, France. – A.-B. Dufour, Lab. de Biométrie et Biologie Evolutive (UMR 5558), CNRS, Univ. de Lyon, Univ. Lyon 1, 43 Bd du 11 novembre 1918, FR–69622 Villeurbanne Cedex, France. – S. Gachet, Muséum National d'Histoire Naturelle, Dépt Ecologie et Gestion de la Biodiversité, 57 rue Cuvier, FR–75005 Paris, France, and IMEP, Univ. Paul Cézanne, FR–13397 Marseille, France.

Functional diversity is at the heart of current research in the field of conservation biology. Most of the indices that measure diversity depend on variables that have various statistical types (e.g. circular, fuzzy, ordinal) and that go through a matrix of distances among species. We show how to compute such distances from a generalization of Gower's distance, which is dedicated to the treatment of mixed data. We prove Gower's distance can be extended to include new types of data. The impact of this generalization is illustrated on a real data set containing 80 plant species and 13 various traits. Gower's distance allows an efficient treatment of missing data and the inclusion of variable weights. An evaluation of the real contribution of each variable to the mixed distance is proposed. We conclude that such a generalized index will be crucial for analyzing functional diversity at small and large scales.

The measurement of distances or similarities among groups of organisms has become a critical step in studies of functional ecology. This increase in interest is largely due to the growth in the number of studies tackling the concept of functional diversity in the last decades (Petchey and Gaston 2006) and to the way that functional diversity is measured. Functional traits of organisms, which are phenotypic traits that enable species to function in their ecosystem, have become fundamental entities for understanding ecosystem processes and for predicting the consequences of environmental modifications, especially on a large scale due to global changes. Here, we consider functional diversity as the variety of states that several functional traits possess in natural conditions.

Various methods for measuring functional diversity exist in the literature (reviewed by Petchey and Gaston 2006). The first method distributes species into functional groups (Walker 1992), and measures functional diversity as the number of functional groups in a given community. The Shannon (1948) or Simpson (1949) index can also be applied to the relative abundances of the groups. Others have proposed the sum and the average of distances between species (Walker et al. 1999, Heemsbergen et al. 2004). Petchey and Gaston (2002) suggested the sum of the branches in a dendrogram (coefficient FD), which can be built using the distances between species. Another alternative is Rao's (1982) quadratic entropy, which includes phenotypic distances among species and an estimation of their abundance (Botta-Dukát 2005). A critical step of all of these indices is defining a general measure of distances based on mixed data. Indeed, phenotypic traits must be measured, and depending on the instruments or experts involved, the variables will be either nominal, ordinal, interval or ratio-scale (Anderberg 1973). Moreover, there may be special cases of scale variable types, such as binary, circular and fuzzy. A potentially high number of statistical types of variables must be integrated and a measure flexible enough to apply to any statistical types of variables must be identified.

Several coefficients of distance or similarity have been developed to handle mixed data sets (Estabrook and Rogers 1966, Gower and Legendre 1986, Carranza et al. 1998). We focused on Gower's (1971) general measure of distance because Gower defined the measure in a mathematical framework associated with interesting properties of Euclidean distances. Gower (1971) proposed measuring a general similarity among entities from the following types of variables: quantitative (variables measured on the interval and ratio scale), nominal, and 'dichotomous' (presence/absence variables). Although his paper was directed towards taxonomists, it has impacted a much larger audience. His measure has been used in a variety of fields, including taxonomy, medicine (Kosaki et al. 1996), genetics (Mohammadi and Prasanna 2003), morphometry (Loo

et al. 2001), paleoecology (Elewa 2004) and physics (Ogurtsov et al. 2002). Our research is motivated by the fact that Botta-Dukát (2005) and Podani and Schmera (2006) recently proposed this metric for the measurement of functional diversity.

The aim of the paper is to show how Gower's metric can be extended to include more types of variables encountered in studies of functional diversity and to highlight its properties. We (1) develop an extension dedicated to functional traits, called 'mixed-variables coefficient of distance', to measure the functional distances among species, (2) demonstrate that this extension can be generalized to handle any type of variables, (3) provide a measure of the contribution of each variable to the global distance, (4) provide a panel of possible analyses for measuring and describing functional diversity from Gower's extended metric, (5) illustrate these theoretical presentations using a field study case, and (6) discuss the performance of the method to mix variables in a context of functional diversity measurement.

Mathematical background

Gower's general coefficient of similarity

The general similarity between species i and j is measured by the following equation:

$$S_{ij} = \sum_{k=1}^{n} s_{ijk}\delta_{ijk}w_k \Big/ \sum_{k=1}^{n} \delta_{ijk}w_k \qquad (1)$$

where n is the number of variables, s_{ijk} is the similarity between i and j calculated on the kth variable, δ_{ijk} is equal to 0 if the value of the kth variable is missing for one of the two species i and j and 1 if it is available for both species, and w_k are the variable weights. According to this equation, the similarity for many variables is a weighted average of similarities for all of the variables that are available for the two species. For each pair of species, the average distance is calculated for a subset of available variables. The values of S_{ij} lie in the interval [0, 1]. The following equation can be used to calculate a coefficient of distance from S_{ij}: $D_{ij} = \sqrt{1 - S_{ij}}$. Gower demonstrated that, without missing data, the matrix $[D_{ij} = \sqrt{1 - S_{ij}}]$ obtained by pairwise comparison is associated with a cloud of points in a Euclidean space.

The first types of variables treated by Gower are measured on the interval and ratio scale. Among various existing metrics, Gower chose the Manhattan metric that calculates the average absolute difference among pairs of values. To normalize the variables, he suggested dividing values by their range (maximum minus minimum values), because the range is easy to calculate and the standard deviation has little meaning for the heterogeneous populations where similarity or dissimilarity coefficients are employed. Let X_k be a variable measured on interval or ratio scale, where parameter k denote the index of the variable out of the n variables considered in Gower's coefficient. Let x_{ik} be the value taken by this variable for species i. Let R_k be the range of X_k either calculated on the observed sample or on the whole population. Let $z_{ik} = x_{ik}/R_k$, for the kth variable X_k, $s_{ijk} = 1 - |z_{ik} - z_{jk}|$. If n variables are used, then Gower's coefficient of similarity is equal to Cain and Harrison's (1958) taxonomic similarity: $s_{ij} = 1 - \Sigma_{k=1}^{n}|z_{ik} - z_{jk}|/n$. Thus, the distance proposed by Gower is the following equation:

$$d_{ij} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}|z_{ik} - z_{jk}|} \qquad (2)$$

Alternatives exist, for example the Euclidean metric:

$$d_{ij} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(z_{ik} - z_{jk})^2} \qquad (3)$$

Gower also distinguished 'dichotomous' variables, which are binary variables with only two levels: 1 (presence) and 0 (absence). Let X_k be a dichotomous variable and x_{ik} be the value taken for this variable for species i. In that case, $s_{ijk} = 1$ if $x_{ik} = 1$ and $x_{jk} = 1$ and $s_{ijk} = 0$ if either x_{ik} or x_{jk} equals zero.

For a nominal variable (X_k), the value for species i are denoted by x_{ik}. $s_{ijk} = 0$ if species i and j disagree in the kth character ($x_{ik} \neq x_{jk}$) and 1 if they agree ($x_{ik} = x_{jk}$). Gower distinguished the special case of two-level nominal variables, qualified as 'alternative variables'. We will not make this distinction and refer to them as 'nominal variables'.

Existing extensions of Gower's distance

Gower's distance has been applied to additional types of variables (Williams and Wentz 2008). Here, we propose to review the extensions that could be useful for the measurement of functional diversity, while other extensions are possible.

Ordinal variables

The main extension of Gower's distance accommodates ordinal variables. The difficulty with ordinal variables is that the operations of subtraction, multiplication and division are not interpretable. Another difficulty is that ties appear for partially ranked variables. Affirming these two difficulties, Podani (1999) suggested one coefficient very specific for ordinal variables but not metric, and another one less specific but metric. The metric alternative corresponds to Eq. 2 applied to ranks.

Multichoice nominal variables

Questions were raised about how to treat binary variables when some of them are associated. For example, a bird species can be both granivorous and frugivorous. In that case, the variable 'trophic habit' is encoded with several columns that are labeled by the trophic states (granivorous, frugivorous). The ith row for species i contains a 1 for each food category it usually uses and 0 elsewhere. These variables can be named 'multichoice nominal variables' in reference to multichoice questions in the sample survey. Podani and Schmera (2007), who tackled this problem explicitly in the context of Gower's formula, used the expression 'trait with non-exclusive states'. Numerous coefficients of distance have been proposed for multichoice nominal variables, such as the simple matching coefficient

or the complement of Jaccard's coefficient (Gordon 1990, reviewed by Legendre and Legendre 1998).

Methods

Toward a more general index of functional distances

Gower's coefficient as the mean of squared distances

Extensions of Gower's coefficient are possible; however, such extensions or merely such possibilities of extensions are scattered in literature. They are scarcely known and, as far as we know, have never been clarified into a general framework. There is a pressing need for a synthesis of the extensions of Gower's distance because these extensions can be used in the framework of functional diversity. Indeed, Botta-Dukát (2005) concluded his article by writing the following: "If categorical and qualitative traits are considered in the same analysis, the number of potential distance functions is strongly limited. Development of a new function more flexible than the Gower distance would be effective."

Gower's distance formula is as follows:

$$D_{ij} = \sqrt{1 - \sum_{k=1}^{n} s_{ijk}\delta_{ijk}w_k \Big/ \sum_{k=1}^{n} \delta_{ijk}w_k} \qquad (4)$$

The possibility to weight variables through the w_k values is useful in functional ecology because "if in reality some traits are more important for determining ecosystem functioning than others then they should be given greater weighting in the trait matrix" (Petchey and Gaston 2002). Let us introduce $d_{ijk} = \sqrt{1\text{-}s_{ijk}}$. We will now complete the discussion in terms of distances instead of similarities.

$$D_{ij} = \sqrt{\sum_{k=1}^{n}(1 - s_{ijk})\delta_{ijk}w_k \Big/ \sum_{k=1}^{n} \delta_{ijk}w_k}$$
$$= \sqrt{\sum_{k=1}^{n} d_{ijk}^2\delta_{ijk}w_k \Big/ \sum_{k=1}^{n} \delta_{ijk}w_k} \qquad (5)$$

Consequently, for many variables, the global distance between two species is the squared root of the average squared distances between species for all the variables considered. Let $\Delta_k = [d_{ijk}]$ be the matrix of pairwise distances between species for the kth variable, and let $\Delta_{mean} = [D_{ij}]$ be the average matrix of pairwise distances between species. If for all k, Δ_k is Euclidean, then Δ_{mean} is Euclidean, even if the values in Δ_{mean} are not comprised between 0 and 1. The Euclidean property is assured by (1) the fact that each function used on a variable (whatever its type) is a metric with Euclidean properties and (2) the use of a weighted mean on the squared distances, instead of the raw distances (demonstration in Appendix 1). We call D_{ij} the 'mixed-variables coefficient of distance'.

Including more types of variables

Gower distance is actually flexible. In this section, we explain how circular and proportion variables can be included in the mixed-variables coefficient of distance. These types of variables are very useful when measuring phenotypic traits in a view of capturing a functional diversity. For example, they allow seasonal traits to be circular variables and diet habits in animals and dispersal mode in plants to be fuzzy variables.

Podani and Schmera (2006) stated that one variable had to be corrected for circularity in their case study, but they did not explain how that transformation was done. There are, however, existing formulas that handle circular variables. For example, Jammalamadaka and SenGupta (2001) presented the following two distances:

$$d_0(\alpha, \beta) = \min(\alpha - \beta, 2\pi - (\alpha - \beta))$$
$$= \pi - |\pi - |\alpha - \beta|| \qquad (6)$$

$$d(\alpha, \beta) = 1 - \cos(\alpha - \beta) \qquad (7)$$

where α and β are given in Fig. 1A. The distance d_0 lies in $[0,\pi]$, while d lies in $[0, 2]$. Consequently, choosing either d_0/π or d/2 would lead to the desired property of having a distance lying in $[0, 1]$. In addition, the variables used for functional ecology are often evenly distributed on the circle and have a finite possible number of levels. For example, there are 12 months in a year (we do not know exactly when in each month the event happened) and four seasons in a year (Fig. 1B). An evenly distributed circular variable is composed of m levels, which are numbered from 1 to m. All of the levels are not necessarily present in the data set. We define below a measure for treating such variables based on

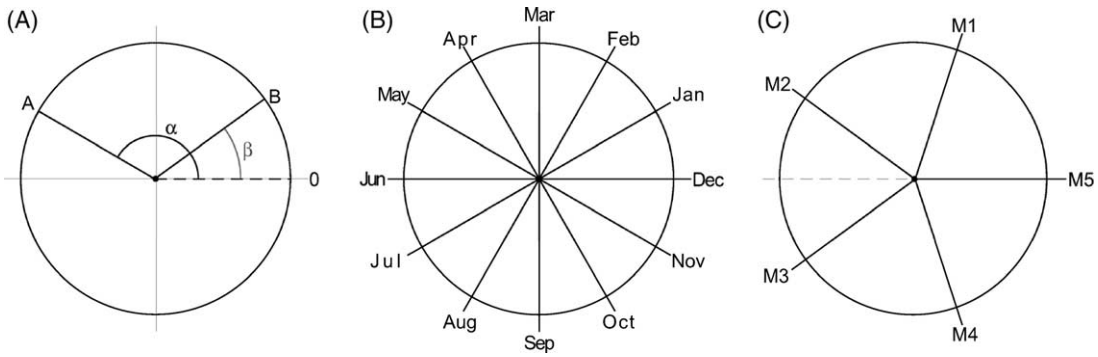

Figure 1. Circular variables. (A) circular distances are often defined as functions of differences in angles. (B) in functional ecology, variables that describe time periodicity are often used. They are characterized by an even distribution along the circle. (C) in case of a finite odd number of levels evenly distributed along the circle (say five levels M1, M2, ..., M5), the maximum theoretical distance (distance between 0 and π) will never be observed. This last property led us to define Eq. 8.

a modification of d_0/π, including a correction for the odd numbers of levels (Fig. 1C). Let X_k be an evenly distributed circular variable, and let x_{ik} be the number of the level taken by species i,

$$d_{ijk} = \begin{cases} \sqrt{1-\left|1-2\left|\frac{x_{ik}}{m}-\frac{x_{jk}}{m}\right|\right|}, & \text{if m is even} \\ \sqrt{1-\left|1-\frac{2m}{m-1}\left|\frac{x_{ik}}{m}-\frac{x_{jk}}{m}\right|\right|}, & \text{if m is odd} \end{cases} \quad (8)$$

The advantage of Eq. 8 is that it provides a Euclidean matrix of values varying from 0 to 1, inclusive. For example in Fig. 1B, if the months are coded by 1 (January) to 12 (December), then the maximal distance is a time lag of six months. The distance between November and May is $\sqrt{1-\left|1-2\left|\frac{11}{12}-\frac{5}{12}\right|\right|} = \sqrt{1-\left|1-\frac{6}{6}\right|} = 1$ while in Fig. 1C, the distance between M1 and M3 is $\sqrt{1-\left|1-2\frac{5}{4}\left|\frac{1}{5}-\frac{3}{5}\right|\right|} = \sqrt{1-\left|1-\frac{4}{4}\right|} = 1$

Consider again the example of feeding habits. Suppose that we have a more detailed idea of the affinity of a species for each feeding category. For example, if we defined a macroinvertebrate species behavior as shedder, scraper and engulfer, do we know whether it spends more time being a shedder, scraper, engulfer? Thus, the affinity can be measured as the proportion of time spent at each activity. It can also be measured according to a fuzzy coding scheme if the determination of the affinity is provided by the global knowledge of an expert, instead of by an experimental measurement. Therefore, affinities are rarely precise; instead, they provide a 'best we can do' attitude as for the treatment of missing data (Estabrook and Rogers 1966). The affinity for a level lies from no affinity (0) to high affinity (fixed to a number specified by the expert). Let a_{imk} be the affinity of species i for the level m of the kth variable, $1 \le m \le M_k$. Fuzzy variables can be transformed into proportion variables via $q_{imk} = a_{imk}/\Sigma_m a_{imk}$ (Chevenet et al. 1994, Bady et al. 2005). Let X_k be a variable defined on $P = \{(p_1, \ldots, p_m, \ldots, p_{M_k}) | \Sigma_{m=1}^{M_k} p_m = 1,\ p_m \ge 0\}$. The value taken by species i is the vector $(q_{i1k}, \ldots, q_{imk}, \ldots, q_{iM_kk})$. As for multichoice nominal variables, numerous distance metrics have been suggested to treat variables that are expressed as proportions of several levels (Legendre and Legendre 1998).

The choice of each metric for each type of variable should be justified with both statistical and biological arguments. In the first case, one might justify their choice by affirming that the selected metric will subsequently improve statistical methods that will be applied to the distances. For example, Milligan and Cooper (1988) found that the standardization by the range for interval and ratio scale variables improved the step of classification methods. They concluded with "Deciding on a suitable form of standardization of variables can improve recovery of the true cluster structure, but it is only one of the several decisions faced by the applied researcher". Several metrics have been developed in the context of a precise application, such as niche recovery, which will help users to decide.

Measuring the contribution of each variable to the global distance

Even if the weights (w_k) of the variables in the calculation of the global distance are equal, the contributions of the variables can be different. Let $\mathbf{d}_k$ be the vector with the S(S–1)/2 pairwise distances between species for the kth variable, where S is the number of species. Without missing data, the correlation between the squared pairwise distances defined by the kth variable and the global squared distances defined by the mixed-variables coefficient of distance is equal to

$$\mathrm{cor}\left(\mathbf{d}_k^2, \sum_{l=1}^{n} w_l \mathbf{d}_l^2\right) = \frac{\sum_{l=1}^{n} (w_l \sqrt{\mathrm{var}(\mathbf{d}_l^2)})\mathrm{cor}(\mathbf{d}_k^2, \mathbf{d}_l^2)}{\sqrt{\mathrm{var}\left(\sum_{l=1}^{n} w_l \mathbf{d}_l^2\right)}} \quad (9)$$

The term $\sqrt{\mathrm{var}(\Sigma_{l=1}^{n} w_l \mathbf{d}_l^2)}$ is positive and does not influence the relative contribution of $\mathbf{d}_k^2$ to the global squared distance. Consequently, even if $w_l = 1/n$ for all l, the relative contribution of $\mathbf{d}_k^2$ to the squared global distance will be higher if it has high correlation with the squared distances obtained on the other variables that lead to the highest variance of squared distances. These correlation values inform thus on the contribution of each variable to the global distance. For n variables verifying, $\mathrm{cov}(\mathbf{d}_k^2, \mathbf{d}_l^2) = 0$ for all $k \neq l$, the contribution of a variable in the global distance will still depend on its weight w_k and the variance of the squared distances obtained from it ($\mathrm{var}(\mathbf{d}_k^2)$):

$$\mathrm{cor}\left(\mathbf{d}_k^2, \sum_{l=1}^{n} w_l \mathbf{d}_l^2\right) = \frac{\sqrt{w_k^2 \mathrm{var}(\mathbf{d}_k^2)}}{\sqrt{\left(\sum_{l=1}^{n} w_l^2 \mathrm{var}(\mathbf{d}_l^2)\right)}} \quad (10)$$

If those variances are equal ($\mathrm{var}(\mathbf{d}_k^2) = \mathrm{var}(\mathbf{d}_l^2)$) for all k, l, then
$\mathrm{cor}(\mathbf{d}_k^2, \Sigma_{l=1}^{n} w_l \mathbf{d}_l^2) = \sqrt{w_k^2 / \Sigma_{l=1}^{n} w_l^2}$ and
$\mathrm{cor}(\mathbf{d}_k^2, \Sigma_{l=1}^{n} \mathbf{d}_l^2/n) = \sqrt{1/n}$

Visualizing the distances

Ordination and clustering methods can be used for visualizing distances. Euclidean distances can be embedded in a Euclidean space where the geometric distances between points are exactly equal to the focus distances. This representation is obtained by principal coordinate analysis (PCoA) (Gower 1966). Each axis in the PCoA maximizes the statistic $\Sigma_{i=1}^{s} \Sigma_{j=1}^{s} \frac{1}{S}\frac{1}{S}\frac{d_{ij}^2}{2}$

This statistic, which can be considered as a measure of functional diversity for the global data set, is equal to the average half-squared distance between species. If the distances are not Euclidean, then PCoA provides a distorted scatter of points with dimensions in an imaginary space. If the absolute sum of negative eigenvalues is low, then PCoA can still be considered. Otherwise, transformations (Lingoes 1971) or non-metric multidimensional scaling (Kruskal

1964) are useful alternatives. More complicated methods may be envisaged depending on the objective of the study; for example, discriminant analyses based on distances can be used if species have to be included into groups (Arenas and Cuadras 2002). Pavoine et al. (2004) developed a double principal coordinate analysis (DPCoA), which measures diversity by Rao's (1982) quadratic entropy and provides a graphical description of the diversity within and between sample units.

Several clustering processes have been developed. Podani and Schmera (2006) suggested the use of the average link (UPGMA) for Gower's distance (but see Petchey and Gaston 2007 for a critical comment). The distances among species that are calculated by the sum of branches in the minimum path connecting them on the dendrogram are ultrametric. A $n \times n$ matrix $\mathbf{D} = [d_{ij}]$ is ultrametric if and only if $d_{ij} \geq 0$, for all i and j, $d_{ij} \leq \max(d_{ik}, d_{kj})$, for all i, j and k, and $d_{ii} < \min_{j \neq i}(d_{ij})$, for all j ($d_{ii} = 0$). Therefore, the dendrogram can be seen as an 'ultrametric representation of a dissimilarity matrix'. Processes have been developed to find the ultrametric that minimizes the least square distance to a given distance matrix (de Soete 1986); we suggest that this procedure could be a relevant alternative to the more well-known clustering analyses for measuring functional diversity.

Studying diversity from the distances

For measuring functional diversity, several indices can include phenotypic differences among species. Ordination analyses, or more often clustering methods, can serve to design functional groups of species. These analyses have been very frequent over the last decades, but are now controversial largely because within-group diversity is eliminated. The functional groups and the functional diversity index (FD) (Petchey and Gaston 2002) depend on the quality of the clustering method selected. On the other hand, the average distance and the sum of all pairwise distances avoid using clustering methods. Regarding Rao's (1982) quadratic entropy, Pavoine et al. (2005) demonstrated interesting properties of Rao's index when the distances are ultrametric and the ultrametric property is generally obtained via clustering methods. In that context, the processes that provide the ultrametric minimizing least square distance to a given dissimilarity matrix might be useful. More studies on the impact of clustering methods on the measurement of functional diversity are necessary.

A case study

We programmed a flexible function for R (R Development Core Team 2007) available in Supplementary material Appendix 1, with a manual in Supplementary material Appendix 2. It can handle interval, ratio scale, dichotomous, nominal, ordinal, circular, multichoice nominal and fuzzy variables.

We analyzed a data set of 80 plant species collected in 15 periurban woodlands with a total of 75 quadrats and species were characterized by 13 phenotypic variables of differing types (Table 1). This data set is described in Appendix 2 and Supplementary material Appendix 3–5. Ratio scale variables are treated by Euclidean metric (Eq. 3). Nominal variables are treated as in Gower (1971). The circular variable is treated by Eq. 8, and ordinal variables are treated by Eq. 3 applied to ranks. Justifications for all of these metrics have been given in previous sections. To compute dissimilarities from the fuzzy variables, we selected the Orloci's chord distance (Orloci 1967) defined as

$$D_{ijk} = \sqrt{2}\sqrt{1 - \sum_{m=1}^{M_k} q_{imk} q_{jmk} \Big/ \sqrt{\left\{\sum_{m=1}^{M_k} [q_{imk}]^2 \sum_{m=1}^{M_k} [q_{jmk}]^2\right\}}} \quad (11)$$

where q_{imk} and q_{jmk} are the percentage of affinity of species i and j, respectively, for the level m of the kth variable.

Table 1 Variables used for the description of plants

Code	Variable	Statistical type	Description
li	ligneous[1]	nominal	presence or absence of ligneous structures
pr	prickly[1]	nominal	presence or absence of prickly structures
fo	start month of flowering period[2]	circular	month when the flowering period starts
he	plant height[2]	ordinal	maximum height of the leaf canopy (from 1: <10 cm to 8: >15 m)
ae	aerial vegetative multiplication[2]	ordinal	from 0: lack of aerial vegetative multiplication; 1: vegetative multiplication occurring infrequently or only on very short distances to 2: vegetative multiplication occurring frequently
un	underground vegetative multiplication[2]	ordinal	same scale as aerial vegetative multiplication
lp	leaf position[2]	nominal	rosette, semi-rosette (rosette before the flowering period), leafy stem
le	leaf persistence[2]	nominal	leaves: seasonal aestival; seasonal hibernal; seasonal vernal; always evergreen; partially evergreen
mp	mode of pollination[3]	fuzzy	respective frequency of autopollination, pollination by insects and pollination by wind
pe	life-cycle[2]	fuzzy	respective frequency of annual, monocarpic (but live more than one year) and polycarpic life cycles
di	dispersion[2]	fuzzy	respective frequency of dispersion by ants; ingestion by animal; external transport by animals; transport by wind; unspecialized transport
lo	seed bank longevity index[4]	ratio scale	index proposed by Bekker et al. [5] in order to take into account results obtained from different studies. The index ranges from 0 (strictly transient) to 1 (strictly persistent)
lf	length of flowering period[2]	ratio scale	number of months of the flowering period

[1] Field observations by J. Vallet, [2] (Grime et al. 1988), [3] (Kühn et al. 2004), [4] (Thompson et al. 1997), [5] (Bekker et al. 1998)

Its value lies in $[0, \sqrt{2}]$. To obtain a metric with Euclidean properties that is bounded between 0 and 1, we used

$$d_{ijk} = D_{ijk}/\sqrt{2} \quad (12)$$

First, we calculated the representation of each variable in the global distance by using Eq. 9. We then applied a PCoA using Lingoes (1971) transformation to render our distance matrix Euclidean. The particular variables considered in this paper are circular and fuzzy. Consequently, we applied the PCoA to a circular variable (start month of flowering period) and a fuzzy variable (mode of pollination), separately. We used Eq. 8 to compute distances between species based on the start month of flowering period, and we used Eq. 12 to compute distances between species based on the mode of pollination. As indicated in the text, we also transformed the distances into ultrametric distances by minimizing the least square difference between the raw distances and the transformed ultrametric distances (de Soete 1986). These ultrametric distances were used for calculating functional diversity within quadrat by two indices: Petchey and Gaston (2002) FD index and the average distance between pairwise species. Two additional diversity indices were included: the species richness within the quadrats and the equitability between ligneous and herbaceous species measured by four times the product of the proportion of herbaceous species and the proportion of ligneous species (index lying between 0 and 1). We used DPCoA to describe diversity within and between quadrats and a principal component analysis (PCA) to compare the four diversity indices.

Results

Most of the correlations among the squared distances obtained by pairs of variables were close to zero, with a mean of 0.036 and a standard deviation of 0.111, suggesting low redundancy between the variables (Fig. 2A). However, five variables have higher correlations with each other and are consequently well represented in the final distance (Fig. 2B): ligneous versus herbaceous, mode of dispersion, plant height, leaf position and leaf persistence.

The first axis of the PCoA (Fig. 3) mainly separates ligneous species that use endozoochory as a way of dispersal from herbaceous species, that include species with rosettes and semi-rosettes and mostly use epizoochory. For the circular variable itself, a perfect symmetrical figure, like the circle displayed in Fig. 1B, would be obtained if all of the months were represented with equal frequencies in the data set (i.e. an equal number of species having each start month for the flowering period). For our data set, however, each month was not represented with equal frequencies; therefore, we obtained Fig. 4A and 4C. The cloud of points is included in a six-dimensional space. On the first two axes, which express 50% and 23% of the average half-squared distance between species, points form a curve starting from January to September. The fuzzy variable describes three levels (autopollination, pollination by insects and pollination by wind). On the first two axes of the PCoA (Fig. 4D), which express 79% and 20% of the average half-squared distance between species (Fig. 4B), the convex hull enclosing the points looks like a slightly distorted triangle. Species that are specialized for one of the three modes of pollination are located on the vertices of the triangle-like hull. Species that use two modes of pollination are located on the edges of the triangle-like hull. Their exact position depends on the affinity of the species (expressed as percentage) for each of the two modes. For example, *Stellaria holostea* is located at x = 0.27 and y = 0.07 on Fig. 4D. Its affinity for autopollination is 25% and for insect pollination is 75%. Therefore, this species is located on the edge of the triangle-like hull that connects species that use autopollination with species that are specialized for pollination by insects. It is closer to insect pollinated species because it has a higher affinity for this mode of pollination. No species in our data set had positive affinities for more than two modes of pollination. These species would have been located inside the triangle-like hull. This reasoning is also valid for more complicated convex hull structures that have more than three levels within the fuzzy variable.

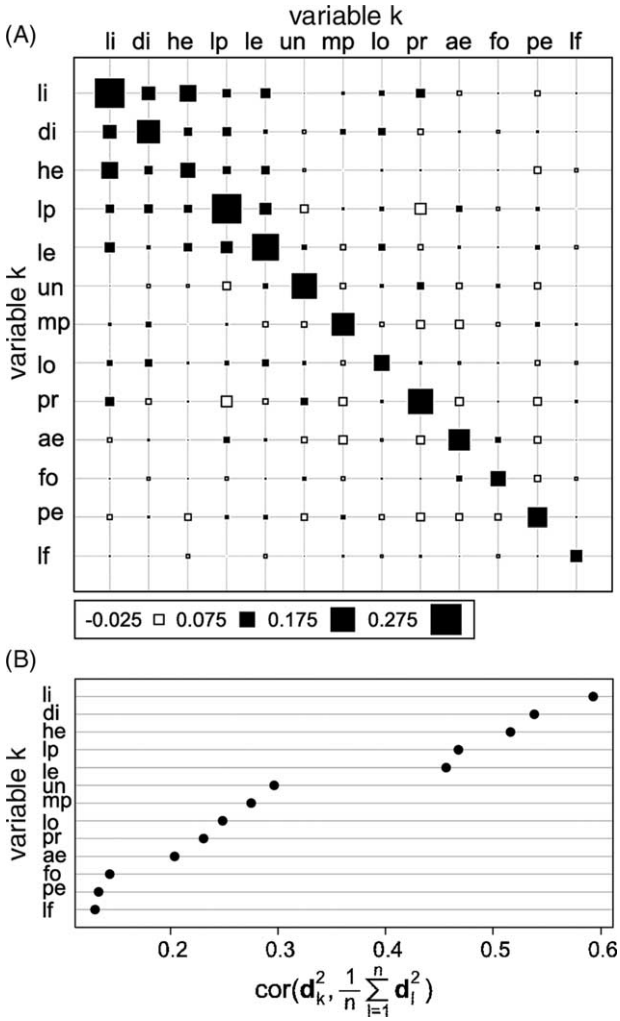

Figure 2. Covariances, variances and contributions of squared distances obtained from variables: (A) variance/covariance matrix between the squared distances obtained for pairwise variables; (B) contribution of each variable to the global distance obtained by Eq. 9 and displayed by a Cleveland's (1994) dot plot.

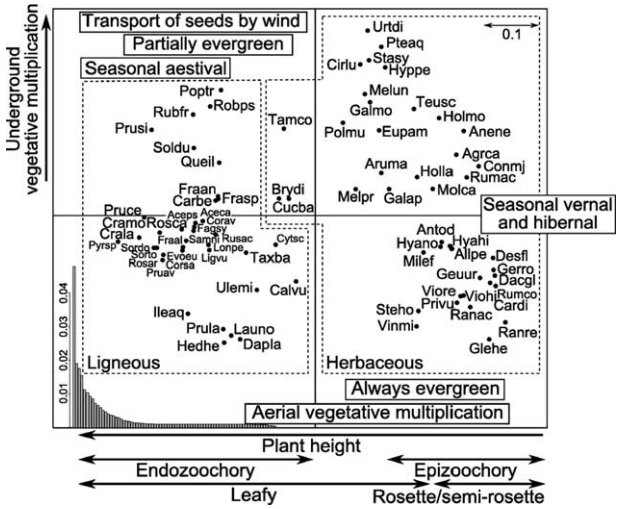

Figure 3. Principal coordinate analysis (PCoA) applied to the global distances among species. Axis 1 (horizontal) expresses 20% of the variation and axis 2 (vertical) expresses 8% of the variation. The eigenvalue barplot is provided at the bottom left-hand corner of the factorial map. The labels of the species are given by codes; full Latin names are given in Appendix 2. Some of the variables or levels of variables have been added to the graph to help the interpretation. The most clear-cut variable (ligneous versus herbaceous) is displayed by grouping species in dashed boxed. The variables located outside the map are associated with arrows indicating clear separations of the species according to the first or second axis, and the framed variables located inside the map indicate tendencies. The scale is given in the top right-hand corner of the map.

The first axis of the DPCoA (Fig. 5A) expresses 42% of the diversity between quadrats. It is highly correlated ($r = 0.93$) with the first axis of the PCoA applied to the same distances between species (abscissa axis in Fig. 3). According to Fig. 5C–5F, the less diverse quadrats in term of average functional distance, which are on the left of Fig. 5B, contain only or mostly ligneous plants, especially the two most common species, *Hedera helix* and *Rubus fruticosus*. In addition to ligneous species, the most diverse quadrats also contain abundant herbaceous species that are rarely found in the other quadrats, including *Alliaria petiolata*, *Dactylis glomerata*, *Geranium robertianum*, *Holcus mollis*, *Hyacinthoides non-scripta* and *Stellaria holostea*, depending on the quadrat. The addition of these herbaceous species increased the functional diversity of the quadrats as shown in Fig. 5C–D, mostly in term of average distance. The correlations between the first axis of the DPCoA (Fig. 1) and the four diversity indices are 0.79 (equitability between ligneous and herbaceous species), 0.73 (average distance), 0.57 (FD), and 0.45 (species richness). A total of 13 quadrats only contained ligneous species. The average global functional distance within those quadrats was lower while the index FD depended on the number of species in each quadrat (Fig. 5G–H). The other quadrats containing also herbaceous species generally displayed higher average functional distance between species especially if the balance between herbaceous and ligneous species was even, but not necessarily higher FD values. However even if the difference between ligneous and herbaceous is the main factor expressed in the global distances, species richness and functional differences within the group of ligneous species and within the group of herbaceous species also influenced the values of functional diversity within quadrats. For example, the point highlighted by a star in Fig. 5G corresponds to a quadrat (BAN5 see codes in Supplementary material Appendix 4) with only nine species (against 18 for the richest quadrat) and no herbaceous species but displays a high FD value. The nine ligneous species it contains are characterized by a large range of vegetative as well as reproductive trait values.

Discussion

The mixed-variables coefficient of distance is a simple index that corresponds to the squared root of the average squared distance between species over all the variables considered. We presented ways of standardizing the variables between 0 and 1, so that the distances obtained from different statistical types of variables will not be skewed by differences of scales. The Eq. 9 provides a solution to evaluate the relative contribution of each variable to the global distances obtained from the mixed-variables coefficient of distances. Those contributions can differ even if equal weights are given to the variables. The global distances were used to obtain a description of the functional diversity within and between species assemblages. Here we discuss the performance of the method to mix variables, to provide details on the result of the mixing, and to improve measurement of functional diversity.

Performance of the method

Botta-Dukát (2005) stated that the following questions must be considered when choosing a measure of distance: (1) in which scale are the traits measured? (2) is standardization of character values desirable or not? (3) is log-transformation of character values possible and meaningful or not? (4) are correlations among descriptors taken into account? Concerning the first two questions, all traits are standardized in the mixed-variables coefficient of distance so as to obtain distances between 0 and 1 for each trait and for the global distance. For ratio-scale variables, the standardization by the range assures that the resulting distance is not modified by a change in the scale of measurement (for example cm or m for tree heights). Concerning the third question, log-transformation of ratio-scale variables is possible, and useful in case of skewed distribution of the values to avoid a high effect of extreme values. As for the fourth question, the correlations between the variables are not removed in the global distance, but their effect on the contribution of each variable to the global distance can be calculated. If only ratio-scale variables are considered, a well-known metric of distance that removes the correlations between variables is the Mahalanobis distance. As far as we know, no equivalent metric exist when mixed variables are considered.

Relative contributions of the variables

Even if the distances are constrained between 0 and 1 and the values of w_k are equal for all k, the contributions of the

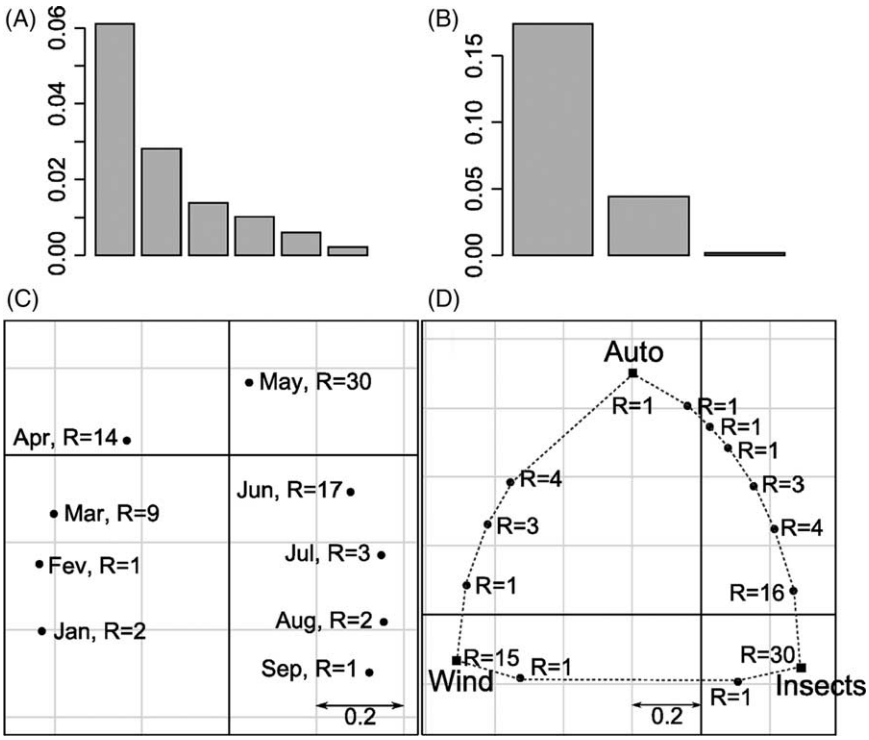

Figure 4. Principal coordinates analysis (PCoA) applied to distances on a circular and a fuzzy variable. Figures on the left illustrate the PCoA applied to distances between species based on the start month of flowering period (circular variable) calculated with Eq. 8. (A) Eigenvalue barplot; (C) factorial map, the abscissa is the first axis of the PCoA and the ordinate is the second axis. Figures on the right illustrate the PCoA applied to distances between species based on the mode of pollination (fuzzy variable) calculated with Eq. 12. The spore-bearing fern, *Pteridium aquilinum*, was removed from the analysis of pollination because it lacks pollen. (B) Eigenvalue barplot; (D) factorial map, the abscissa is the first axis of the PCoA and the ordinate is the second axis. In panel (C) and (D), R indicates the number of species clustered on a given location. In panel (C), each point represents the month during which the flowering period started. On panel (D), the squares represent specialized species for a single mode of pollination, which is specified. The broken lines define the convex hull of the scatter of points.

variables included in the global distance can differ. First the alternative choice of sample range versus population range for interval and ratio scale variables might have an overwhelming effect on the distances and therefore the functional diversity. Leps et al. (2006) illustrated this point by considering tree height. If the sample is composed of grasslands and if the range is calculated on this sample, then differences of a few centimeters between the heights of individual meadow species might largely contribute to functional diversity. If the whole community also includes woody vegetation, however, then the height distances between grassland species and consequently the height diversity within meadows become negligible. In this latter case, the weight of the variable height will be low in the calculation of the average functional distances, resulting in a variable with observed distances close to zero.

Next, once the standardizing schemes have been selected, the contributions of the variables in the final mixed distance depend on the correlations among variables. If highly correlated variables are included in the calculation of the functional distance, then the information shared by redundant variables will have an exaggerated weight in the final functional diversity. The representation of a trait in the global distance depends on its correlation with the squared pairwise distances it generates and the squared pairwise distances generated by the other variables, especially those with high variances. Variables leading to high variance of pairwise distances between species will thus be more likely to have a high contribution into the global distance, provided that they have correlations with some other variables. Clear-cut variables such as a nominal variable with two levels that define two groups and equitability of the distribution of species into the two groups lead to high variance in pairwise distances. Such variables are likely to be more influent in the calculation of the global distance.

Studying functional diversity

One of the main advantages of obtaining a mixed-variable coefficient of distance in functional ecology is that current indices of functional diversity are based on distances and numerous studies collect functional traits from various statistical types. We highlighted that such an index can be used in diversity indices based on distances such as the FD index, the average and sum of distances, the quadratic entropy. It can also be associated with clustering and ordination methods. The principal coordinate analysis and the double principal coordinate analysis are linked with the average squared distance and quadratic entropy indices of diversity, while FD is related with clustering methods. Therefore distances, clouds of points in multidimensional

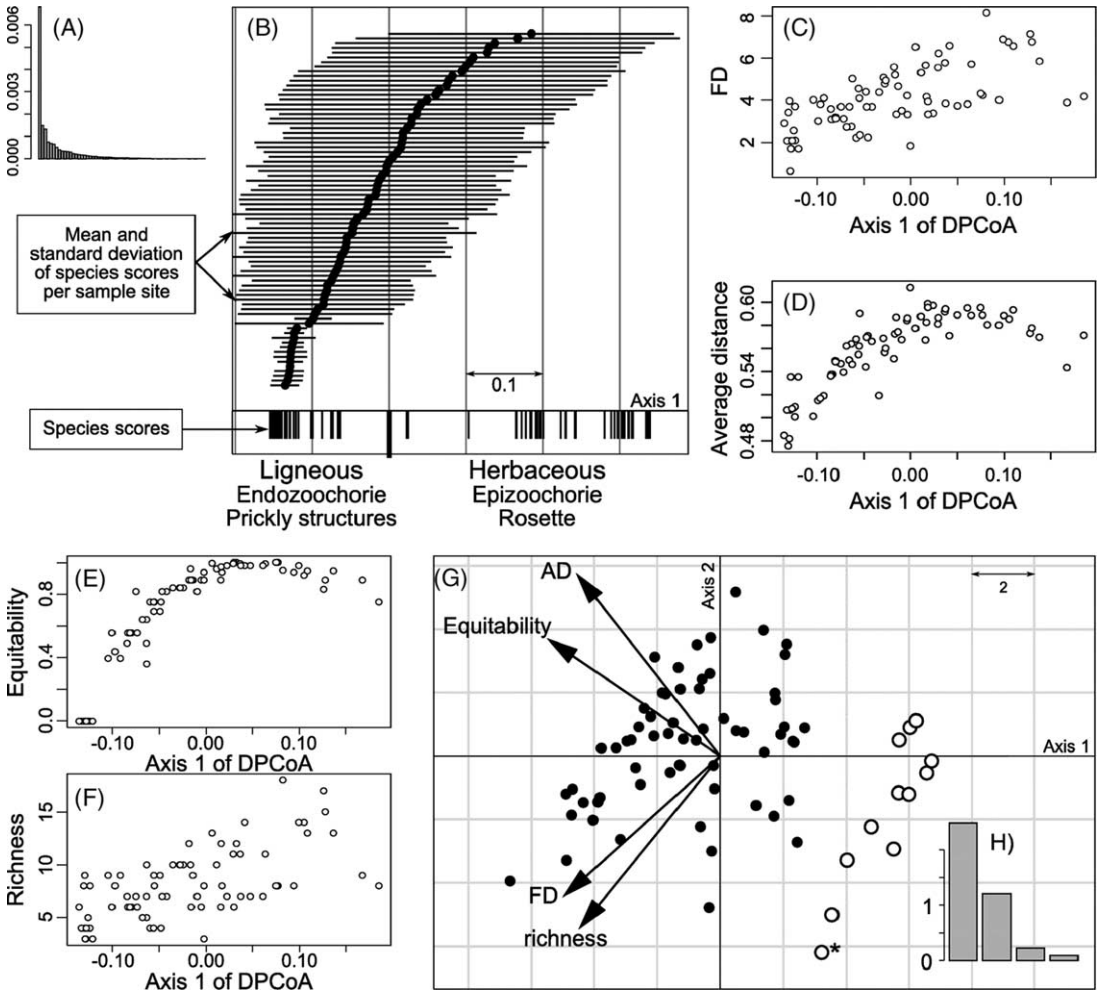

Figure 5. Functional diversity analysis between and within quadrats. We removed one quadrat (FON5, see data set in Supplementary material Appendix 4) because it contained only one species (*Rubus fruticosus*). Panels (A) to (F) are graphical representations associated with the DPCoA applied on the presence/absence of species in quadrats and the ultrametric distances between species: (A) eigenvalue barplot (the first axis expresses 42% of the variation in quadrat points); (B) scores of species and quadrats on the first axis of the DPCoA. Quadrats are located at the average score of the species they contain. The standard deviation of the scores of the species are given for each quadrat; (C) Petchey and Gaston FD index measured for each quadrat as a function of the quadrat scores on the first axis of the DPCoA; (D) average distance between species within quadrats as a function of the quadrat scores on the first axis of the DPCoA; (E) the equitability between ligneous and herbaceous species measured by 4 times the product of the proportion of herbaceous species and the proportion of ligneous species (index between 0 and 1) as a function of the quadrat scores on the first axis of the DPCoA; (F) species richness as a function of the quadrat scores on the first axis of the DPCoA. Panels (G) and (H) analyze the correlations between the four diversity indices: (G) first factorial map of the principal component analysis applied to four variables: the species richness within quadrats, the FD index, the average distance (AD) between pairwise species, and the equitability between ligneous and herbaceous species. The open circles indicate quadrats with only ligneous species. The closed circles denote quadrats with both ligneous and herbaceous species. The star highlights a quadrat with a relatively high FD value despite the complete absence of herbaceous species; (H) the eigenvalue barplot is provided in the bottom right-hand corner.

space and at the tips of ultrametric trees, and functional diversity indices are connected.

Each kind of variables leads to a specific shape of cloud of points in a univariate or multivariate Euclidean space. This will have influences on the value of the functional diversity indices. For example, nominal variables separate species into distinct groups, with equal unit distances between the groups. This leads to regular-polygon shapes where each group is a distinct vertex of the polygon. If used with only one nominal variable, the index FD will be a function of the number of groups represented in the community, while the average distances will depend both on the number of groups represented and the equitability of the distribution of species into the groups. If the relative abundances of the species are used then Rao's quadratic entropy index will depend on the number of groups represented and the equitability of the distribution of individuals into the groups. The circular variables lead to circles, arcs of circles, or arcs of ellipsoids. The proportion variables lead to more or less distorted regular polygons delimiting the subspace within which species points are located. Finally, each ratio-scale and ordinal variable leads

to a continuous one-dimensional cloud of points. The shape of the global distances, which will influence the value of the functional diversity indices (Pavoine et al. 2005), results from the combination of these different structures and from the correlations between the variables, even variables from different statistical types.

Case study

The case study concerned 80 plant species and 13 variables in vegetative and reproductive trait space. The best represented trait in the global distance was the distinction between ligneous and herbaceous species. This trait led to the highest correlations with the distances obtained from other traits and had the highest variance of pairwise squared distances. Despite the plant height led to a low variance in squared distances between species, it was well represented in the global distance, because of its high correlation with the ligneous/herbaceous nominal trait. Three of the four nominal traits were associated to the highest variance in the inferred squared distances between species. This highlights the characteristic of nominal variables as providing clear-cut distinction between species associated with high variance of pairwise distances. Thanks to the matrix of correlation between squared distances, we are able to know what kind of diversity will be measured by using the global distances. Here the diversity mostly increases by the addition of herbaceous species to ligneous species assemblages, but not only, as illustrated by the low correlation between the FD index and the equitability of the distribution of species between ligneous and herbaceous groups. In addition, the other variables provide squared distances that have correlations with the global squared distances varying from to 0.13 to 0.53, and thus influence the measures of functional diversity.

The first axis of Fig. 3 mostly separate ligneous from herbaceous species, but the separation is not perfect. If only the variable ligneous/herbaceous was considered, we would have obtained two points, that would have represented the two groups, on the opposite side of the first axis. Instead of two points, the species are continuously dispersed along the first axis and only 20% of the distances between species are represented by this first axis. Even three herbaceous species are on the left part of the axis with the ligneous species. Those three species (*Bryonia dioica*, *Cucubalus baccifer* and *Tamus communis*) exhibit endozoochory, aestival leaf persistence, and pollination by insects, like most of the ligneous species. The five most contributing variables all influence the distribution of points on this first axis. The sixth most contributing variable (underground vegetative multiplication) is correlated with the second axis.

The functional diversities of the quadrats were different according to the index used. FD was more related to species richness and the AD to the balance between ligneous and herbaceous species. FD measures the extent of complementarity among species in the trait space, while AD measures the average distance in a pair of species. Therefore the shared differences between ligneous and herbaceous species are counted only once in FD while they are counted in each pairwise comparison in the AD index. Because the DPCoA method measures point dispersion by the average squared distances, it is mostly related to the AD index.

Following this exploratory study, a selection of traits might be done by considering the subset of traits correlated with a focus ecosystem process (Petchey and Gaston 2006) or environmental gradient (e.g. difference between edges and centers of woodlands, urbanization gradient). Solutions must also be taken if we want to remove the strong effect of the difference between ligneous and herbaceous species (e.g. separated analyses).

Conclusion

In conclusion, Gower's index allows us to consider various traits, with or without missing data. Extending Gower's measure with a more wide range of variable types will enable us to compute distances among species at the various scales where the functions of the species are studied. Whether we can obtain a distance metric that is based on mixed variables and that corrects for the correlations between variables is still an open question. Associating the mixed variables coefficient of distances with diversity indices and ordination methods allows a description of both within and between-sample diversity. Moreover, it improves our possibilities for including various variables when comparing functional diversity to measurements derived from species count and phylogeny. A wide range of applications is possible, because the use of functional distances between species in studies of community structures and dynamics is increasing.

Acknowledgements – We thank Owen Petchey for constructive comments on this paper. This work was funded by a Young Researcher Award obtained from the French Institute of Biodiversity in 2004 and partly supported by Agence Nationale de la Recherche grant ANR-06-JCJC-0032-01. The authors are also very grateful to the "Conseil Général de Maine-et-Loire" for its financial support in collecting field data. SP is supported by the European Commission under the Marie Curie Programme.

References

Anderberg, M. R. 1973. Cluster analysis for applications. – Academic Press.

Arenas, C. and Cuadras, C. M. 2002. Recent statistical methods based on distances. – Contrib. Sci. 2: 183–191.

Bady, P. et al. 2005. Use of invertebrate traits for the biomonitoring of European large rivers: the effect of sampling effort on genus richness and functional diversity. – Freshwater Biol. 50: 159–173.

Bekker, R. M. et al. 1998. Seed size, shape and vertical distribution in the soil: indicators of seed longevity. – Funct. Ecol. 12: 834–842.

Botta-Dukát, Z. 2005. Rao's quadratic entropy as a measure of functional diversity based on multiple traits. – J. Veg. Sci. 16: 533–540.

Braud, S. and Hunhammar, S. 1999. Les ptéridophytes du Maine-et-Loire – inventaire et cartographie. – ERICA 12: 1–61.

Cain, A. J. and Harrison, G. A. 1958. An analysis of the taxonomist's judgement of affinity. – Proc. Zool. Soc. Lond. 131: 85–98.

Carranza, L. et al. 1998. Analysis of vegetation structural diversity by Burnaby's similarity index. – Plant Ecol. 138: 77–87.
Chevenet, F. et al. 1994. A fuzzy coding approach for the analysis of long-term ecological data. – Freshwater Biol. 31: 295–309.
Cleveland, W. S. 1994. The elements of graphing data. – AT and T Bell Laboratories.
de Soete, G. 1986. A least squares algorithm for fitting an ultrametric tree to a dissimilarity matrix. – Pattern Recogn. Lett. 2: 133–137.
Elewa, A. 2004. Quantitative analysis and palaeoecology of Eocene Ostracoda and benthonic foraminifera from Gebel Mokattam, Cairo, Egypt. – Palaeogeogr. Palaeocl. 221: 309–323.
Estabrook, G. F. and Rogers, D. J. 1966. A general method of taxonomic description for a computed similarity measure. – Bioscience 16: 789–793.
Gordon, A. D. 1990. Constructing dissimilarity measures. – J. Classif. 7: 257–269.
Gower, J. C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. – Biometrika 53: 325–338.
Gower, J. C. 1971. A general coefficient of similarity and some of its properties. – Biometrics 27: 857–874.
Gower, J. C. and Legendre, P. 1986. Metric and Euclidean properties of dissimilarity coefficients. – J. Classif. 3: 5–48.
Grime, J. P. et al. 1988. Comparative plant ecology: a functional approach to common British species. – Kluwer Academic Publishers.
Heemsbergen, D. A. et al. 2004. Biodiversity effects on soil processes explained by interspecific functional dissimilarity. – Science 306: 1019.
Jammalamadaka, S. R. and SenGupta, A. 2001. Topics in circular statistics. – World Scientific.
Kosaki, K. et al. 1996. Zimmer phocomelia: delineation by principal coordinate analysis. – Am. J. Med. Genet. 66: 55–59.
Kruskal, J. B. 1964. Nonmetric multidimensional scaling: a numerical method. – Psychometrika 29: 115–129.
Kühn, I. et al. 2004. BiolFlor: a new plant-trait database as a tool for plant invasion ecology. – Div. Distr. 10: 363–365.
Lambinon, J. et al. 1992. Nouvelle Flore de la Belgique, du Grand-Duché du Luxembourg, du Nord de la France et des régions voisines (Ptéridophytes et Spermaphytes). – Jardin Bot. Natl Belgique.
Legendre, P. and Legendre, L. 1998. Numerical ecology. – Elsevier.
Leps, J. et al. 2006. Quantifying and interpreting functional diversity of natural communities: practical considerations matter. – Preslia 78: 481–501.
Lingoes, J. C. 1971. Some boundary conditions for a monotone analysis of symmetric matrices. – Psychometrika 36: 195–203.
Loo, A. H. B. et al. 2001. Intraspecific variation in *Licuala glabra* Griff. (Palmae) in Peninsular Malaysia – a morphometric analysis. – Biol. J. Linn. Soc. 72: 115–128.
Milligan, G. W. and Cooper, M. C. 1988. A study of standardization of variables in cluster analysis. – J. Classif. 5: 181–204.
Mohammadi, S. A. and Prasanna, B. M. 2003. Analysis of genetic diversity in crop plants – salient statistical tools and considerations. – Crop Sci. 43: 1235–1248.
Ogurtsov, M. G. et al. 2002. Long-period cycles of the Sun's activity recorded in direct solar data and proxies. – Solar Phys. 211: 371–394.
Orloci, L. 1967. An agglomerative method for classification of plant communities. – J. Ecol. 55: 193–206.
Pavoine, S. et al. 2004. From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis. – J. Theor. Biol. 228: 523–537.
Pavoine, S. et al. 2005. Measuring diversity from dissimilarities with Rao's quadratic entropy: are any dissimilarity indices suitable? – Theor. Popul. Biol. 67: 231–239.
Petchey, O. L. and Gaston, K. 2002. Functional diversity (FD), species richness and community composition. – Ecol. Lett. 5: 402–411.
Petchey, O. L. and Gaston, K. 2006. Functional diversity: back to basics and looking forward. – Ecol. Lett. 9: 741–758.
Petchey, O. L. and Gaston, K. J. 2007. Dendrograms and measuring functional diversity. – Oikos 116: 1422–1426.
Podani, J. 1999. Extending Gower's general coefficient of similarity to ordinal characters. – Taxon 48: 331–340.
Podani, J. and Schmera, D. 2006. On dendrogram-based measures of functional diversity. – Oikos 115: 179–185.
Podani, J. and Schmera, D. 2007. How should a dendrogram-based measure of functional diversity function? A rejoinder to Petchey and Gaston. – Oikos 116: 1427–1430.
Rao, C. R. 1982. Diversity and dissimilarity coefficients: a unified approach. – Theor. Popul. Biol. 21: 24–43.
Shannon, C. E. 1948. A mathematical theory of communication. – Bell System Tech. 27: 379–423, 623-656.
Simpson, E. H. 1949. Measurement of diversity. – Nature 163: 688.
Thompson, K. et al. 1997. The soil seed banks of north west Europe: methodology, density and longevity. – Cambridge Univ. Press.
Walker, B. H. 1992. Biodiversity and ecological redundancy. – Conserv. Biol. 6: 18–23.
Walker, B. et al. 1999. Plant attribute diversity, resilience, and ecosystem function: the nature and significance of dominant and minor species. – Ecosystems 2: 95–113.
Williams, E. A. and Wentz, E. A. 2008. Pattern analysis based on type, orientation, size and shape. – Geogr. Anal. 40: 97–122.

Supplementary material (available online as Appendix O16668 at <www.oikos.ekol.lu.se/appendix>). Appendix 1. (file: dist.ktab.R) R function for computing the mixed-variables coefficient of distance). Appendix 2. (file: Manual.pdf). Manual for the function (''dist.ktab.R''). Appendix 3. (file: AngersDataSet.pdf) Description of the data set. Appendix 4. (file: flo.txt) Presence/absence of plant species in all quadrats. Appendix 5. (file: traits.txt) Trait values for each plant species.

Appendix 1. Proof that Δ_{mean} is Euclidean

Let $\mathbf{\Delta} = [d_{ij}]$ be a matrix of dissimilarities. $\mathbf{\Delta}$ is Euclidean if and only if, for any vector $\mathbf{a}$ such as $\mathbf{a}^t\mathbf{1} = 0$ then $\mathbf{a}^t\mathbf{D}\mathbf{a} \leq 0$, where $\mathbf{D} = [d_{ij}^2]$.

We imposed that $\mathbf{\Delta}_k = [d_{ijk}]$ be Euclidean, and $\mathbf{D}_k = [d_{ijk}^2]$, which implies that for any vector $\mathbf{a}$ such as $\mathbf{a}^t\mathbf{1} = 0$ then $\mathbf{a}^t\mathbf{D}_k\mathbf{a} \leq 0$. Consider $\mathbf{D}_{mean} = \Sigma_{k=1}^{n}\lambda_k\mathbf{D}_k$, where for any k, λ_k corresponds to $w_k/\Sigma_{k=1}^{n}w_k$, $\lambda \geq 0$, and $\Sigma_{k=1}^{n}\lambda_k = 1$. For all vector $\mathbf{a}$ such as $\mathbf{a}^t\mathbf{1} = 0$, $\mathbf{a}^t\mathbf{D}_{mean}\mathbf{a} = \mathbf{a}^t(\Sigma_{k=1}^{n}\lambda_k\mathbf{D}_k)\mathbf{a} = \Sigma_{k=1}^{n}\lambda_k\mathbf{a}^t\mathbf{D}_k\mathbf{a}$

For all k, $\mathbf{a}^t\mathbf{D}_k\mathbf{a} \leq 0$, consequently, $\Sigma_{k=1}^{n}\lambda_k\mathbf{a}^t\mathbf{D}_k\mathbf{a} \leq 0$, which leads to $\mathbf{a}^t\mathbf{D}_{mean}\mathbf{a} \leq 0$, therefore $\mathbf{\Delta}_{mean}$ is Euclidean. Note that, as Gower (1971) indicated, the Euclidean property is not assured in case of missing data.

Appendix 2. Description of the data set

The study was conducted in Angers conurbation located very close to the Loire River in northwest France (longitude: 00°33′07″W, latitude: 47°28′16″N). It is characterized by an oceanic climate with a mean annual rainfall of 605 mm (Braud and Hunhammar 1999); the monthly mean temperature ranges from 13°C to 24°C in July. The geological substratum is mainly schist. Fifteen woodland stations of around one ha each were surveyed along a rural-urban gradient. Sampling vegetation was undertaken in July 2003. This sampling period permitted the detection of both vernal plants (with dead leaves and fruits) and summer species. These two phenologies are dominant in forest environment. We established five quadrats of 30 m^2 situated in the core area of each woodland. The list of species was established in each quadrat. The nomenclature is taken from Lambinon et al. (1992). Species codes used in Fig. 3 are as follows:

	Species name		Species name
Aceca	*Acer campestre*	Launo	*Laurus nobilis*
Aceps	*Acer pseudoplatanus*	Ligvu	*Ligustrum vulgare*
Agrca	*Agrostis capillaris*	Lonpe	*Lonicera periclymenum*
Allpe	*Alliaria petiolata*	Melpr	*Melampyrum pratense*
Anene	*Anemone nemorosa*	Melun	*Melica uniflora*
Antod	*Anthoxanthum odoratum*	Milef	*Milium effusum*
Aruma	*Arum maculatum*	Molca	*Molinia caerulea*
Brydi	*Bryonia dioica*	Polmu	*Polygonatum multiflorum*
Calvu	*Calluna vulgaris*	Poptr	*Populus tremula*
Carbe	*Carpinus betulus*	Privu	*Primula vulgaris*
Cardi	*Carex divulsa*	Pruav	*Prunus avium*
Cirlu	*Circaea lutetiana*	Pruce	*Prunus cerasifera*
Conmj	*Conopodium majus*	Prula	*Prunus laurocerasus*
Corav	*Corylus avellana*	Prusi	*Prunus spinosa*
Corsa	*Cornus sanguinea*	Pteaq	*Pteridium aquilinum*
Crala	*Crataegus laevigata*	Pyrsp	*Pyrus sp.*
Cramo	*Crataegus monogyna*	Queil	*Quercus ilex*
Cucba	*Cucubalus baccifer*	Ranac	*Ranunculus acris*
Cytsc	*Cytisus scoparius*	Ranre	*Ranunculus repens*
Dacgl	*Dactylis glomerata*	Robps	*Robinia pseudacacia*
Dapla	*Daphne laureola*	Rosar	*Rosa arvensis*
Desfl	*Deschampsia flexuosa*	Rosca	*Rosa canina*
Eupam	*Euphorbia amygdaloides*	Rubfr	*Rubus fruticosus*
Evoeu	*Euonymus europaeus*	Rumac	*Rumex acetosa*
Fagsy	*Fagus sylvatica*	Rumco	*Rumex conglomeratus*
Fraal	*Frangula alnus*	Rusac	*Ruscus aculeatus*
Fraan	*Fraxinus angustifolia*	Samni	*Sambucus nigra*
Frasp	*Fraxinus excelsior*	Soldu	*Solanum dulcamara*
Galap	*Galium aparine*	Sordo	*Sorbus domestica*
Galmo	*Galium mollugo*	Sorto	*Sorbus torminalis*
Gerro	*Geranium robertianum*	Stasy	*Stachys sylvatica*
Geuur	*Geum urbanum*	Steho	*Stellaria holostea*
Glehe	*Glechoma hederacea*	Tamco	*Tamus communis*
Hedhe	*Hedera helix*	Taxba	*Taxus baccata*
Holla	*Holcus lanatus*	Teusc	*Teucrium scorodonia*
Holmo	*Holcus mollis*	Ulemi	*Ulex minor*
Hyahi	*Hyacinthoides hispanica*	Urtdi	*Urtica dioica*
Hyano	*Hyacinthoides non-scripta*	Vinmi	*Vinca minor*
Hyppe	*Hypericum perforatum*	Viohi	*Viola hirta*
Ileaq	*Ilex aquifolium*	Viore	*Viola reichenbachiana*