Journal of Ecology 2007
95, 294–300

FORUM

An operational, additive framework for species diversity partitioning and beta-diversity analysis

RAPHAËL PÉLISSIER and PIERRE COUTERON*
*IRD, UMR AMAP (Botanique et Bioinformatique de l'Architecture des Plantes), TA40/PS2, Bd. de la Lironde, 34398 Montpellier cedex 5, France, and *Institut Français de Pondichéry, Pondicherry, 605001 India*

Summary

1 An important goal of community ecology is the assessment of factors that are likely to influence the spatio-temporal distribution of species assemblages and diversity. Surprisingly, most statistical methods devoted to this have remained poorly interconnected, as well as poorly connected with the popular metrics of diversity estimation. In the present paper we show that important questions related to determinants of species diversity can be specified through a simple multivariate linear model and explored, in common diversity metrics, using standard methods and routines of variance/covariance decomposition.
2 Thanks to an unusual form of presentation of taxonomic data into a *table of species occurrences*, which considers the individuals as data units, Shannon and Simpson indices as well as species richness can all be expressed as a (weighted) sum of squares. Subsequent apportionments into explained and residual sum of squares provide direct estimates of the beta- and alpha-diversity components with respect to either categorical habitat types or continuous gradient variables. Appropriate statistics and non-parametric tests are available to assess the significance of these components.
3 Explicit analytical relationships exist between the linear approximation of the table of species occurrences by sampling sites, and the more classical table of species abundances by sites. Therefore, direct links with methods of ordination in reduced space, such as correspondence analysis and canonical correspondence analysis, provide opportunities for partitions that preserve consistency with usual diversity indices. The sum of squares of the approximated occurrence table provides measures of intersites beta-diversity, from which measures of dissimilarity with explicit references to diversity indices can be derived. Such measures are amenable to distance-based apportionments through multivariate variograms and multiscale ordination.
4 What are the relative effects of the biological, environmental and anthropogenic factors and of their potential interactions on species diversity? Are these effects stable across scales, from landscape to region, between regions and across ecosystems? The methodological integration proposed in our analytical framework enables one to address these questions using standard statistical tools, and opens new prospects for quantitative biodiversity studies. This also paves the way towards refined models for predicting species diversity at unsampled locations.

Key-words: additive diversity partitioning, alpha and beta diversity components, dissimilarity, (generalized) linear model, ordination, spatial scale, species occurrence table, variograms

Journal of Ecology (2007) **95**, 294–300
doi: 10.1111/j.1365-2745.2007.01211.x

Correspondence: Raphaël Pélissier (tel. +33 04 67 61 75 23; fax +33 04 67 61 56 68; e-mail Raphael.Pelissier@mpl.ird.fr).

Introduction

Since the review paper by Lande (1996), there has been a renewed interest in the additive partition of species diversity as a meeting point between theoretical and empirical approaches of community ecology (see References in Veech *et al.* 2002). Indeed, Lande's contribution paved the way to bridging the gap between the concepts of alpha-, beta- and gamma-diversities (Whittaker 1960, 1972) and modern statistical tools. In addition, Lande's paper has stimulated further analytical developments, notably towards scale-dependent apportionments of species diversity and hypotheses testing (e.g. Wagner *et al.* 2000; Crist *et al.* 2003; Kiflawi & Spencer 2004).

However, what has been almost completely ignored is that Lande's approach can also refer to a corpus of standard linear modelling methods largely disseminated amongst ecologists, but generally not used with explicit reference to diversity analysis. We have previously demonstrated (Pélissier *et al.* 2003; Couteron & Pélissier 2004; Couteron & Ollier 2005) that various additive apportionments of species diversity can be achieved within this very general framework, which covers, among others, multivariate analysis of variance, *sensu* Anderson (2001), multivariate multiple regression (including multivariate canonical analysis, *sensu* Legendre & Legendre 1998) and multivariate variography (Wackernagel 1998).

As we have previously focused on particular technical facets, in this paper we illustrate how different aspects of additive diversity partitioning can be assembled into a simple and operational multivariate linear framework that opens opportunities for the joint analysis of and discrimination among different types of processes affecting diversity patterns.

An operational data table

Let us consider a taxonomic relevé in the form of a list of n observations corresponding to a set of individual organisms recorded during a given field survey, and which contains s different species names. The list can be binary-coded as a data matrix with n rows and s columns filled with zeros, except for the unique cell of each row that associates a particular observation with a species name and contains the value 1 (Fig. 1a). Following Gimaret-Carpentier *et al.* (1998), we will call such a matrix a table of species occurrences (an occurrence table, in short). It is noteworthy that any table of species abundances, which originates for instance from the enumeration of s species in a set of p sampling sites and sums across all cells to a total of n observations, can easily be re-coded as a n by s table of species occurrences partitioned according to sites (Fig. 1b), a kind of *inflated data table*, *sensu* Legendre & Legendre (1998, p. 463).

Let us define a hypothetical table of species occurrences, irrespective of sites for the moment, as an $n \times s$ matrix $\mathbf{Y}$ whose element y_{ij} is 1 when the ith observation belongs to species j, 0 otherwise. Total sum of squares of this table is $TSS = \Sigma_{ij}(y_{ij} - y_{\cdot j})^2$, with $y_{\cdot j} = \Sigma_i\, y_{ij}/n$, the relative frequency of species j. The corresponding (biased) variance, i.e. TSS/n, is exactly Simspon index of species diversity (Lande 1996). Introducing a function that modulates weights of species in the above summation, provides diversity quantifications in several popular metrics:

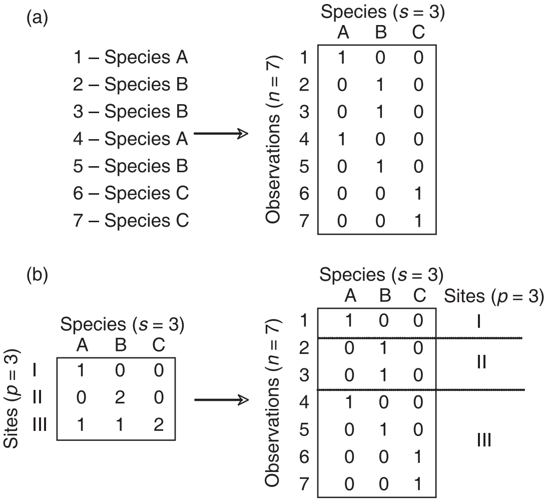


Fig. 1 (a) From a list of $n = 7$ individual observations of $s = 3$ species to a $n \times s$ table of species occurrences, $\mathbf{Y}$. (b) From a table of species abundances of $s = 3$ species in $p = 3$ sites that sums to $n = 7$ observations, to a $n \times s$ table of species occurrences partitioned according to p sites.

$$TSS = \Sigma_j\, w_j\, \Sigma_i (y_{ij} - y_{\cdot j})^2 \qquad \text{eqn 1}$$

Taking $w_j = 1$ for all species, $w_j = \log(1/y_{\cdot j})/(1 - y_{\cdot j})$ or $w_j = 1/y_{\cdot j}$ means equating TSS/n with Simpson diversity, Shannon diversity or species richness (minus one), respectively (Pélissier *et al.* 2003). However, there is in fact no reason to restrict the definition of w_j to functions equating TSS/n with classical measures of species diversity, and one could prefer using weights accounting for the patrimonial, conservation or economic value of species (Yoccoz *et al.* 2001).

An operational multivariate linear model

One of the main goals of community ecology is the identification of environmental factors that are likely to determine the spatial and temporal distribution of species diversity (Gaston & Blackburn 2000). In other words, we would like to be able to quantify the relationship between observed species diversity and one (or a set of) external explanatory variable(s) depicting accessible information about the species' environment. Returning to our above definition of a table of species occurrences, $\mathbf{Y}$, the problem can parsimoniously be specified through the following general multivariate linear model:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \qquad \text{eqn 2a}$$

where $\mathbf{X}$ is a $n \times m$ matrix of explanatory variables, $\mathbf{B}$ a $m \times s$ matrix of unknown parameters, and $\mathbf{E}$ a $n \times s$ matrix of error terms. It could be convenient to specify a model with no intercept by centring the columns of $\mathbf{Y}$ so that their means are all 0 (see Pélissier *et al.* 2003).

How well the model fits to the data means examining how the total variation in table $\mathbf{Y}$ (quantified by *TSS*) partitions into a component explained by predictions of the model or model sum of squares (*MSS*) and a component unexplained by predictions of the model or residual sum of squares (*RSS*). Providing that all the three terms are appropriately weighted via the same w_j function (see previous section), we have $TSS = MSS + RSS$, with:

$$MSS = \sum_j w_j \sum_i (\hat{y}_{ij} - y_{\cdot j})^2 \qquad \text{eqn 2b}$$

$$RSS = \sum_j w_j \sum_i (y_{ij} - \hat{y}_{ij})^2 \qquad \text{eqn 2c}$$

These very general equations hold for any $\mathbf{X}$ matrix, which may contain either quantitative and/or dummy coded qualitative covariates (Sokal & Rohlf 1995; Legendre & Legendre 1998).

Whatever the diversity metric chosen via w_j, the proportion of total species diversity explained by the variables contained in $\mathbf{X}$ can be quantified by the ratio: $R^2 = MSS/TSS = 1 - RSS/TSS$.

A well-known weakness of this ratio is the fact that the denominator is fixed for a given set of observations, while each additional variable in $\mathbf{X}$ can only increase the numerator and thus the R^2 value, even though the new variable is completely random. Moreover, as the model intrinsically aims to predict species identity for a potentially very large number of individual occurrences, *RSS* is inevitably large and the R^2 value is likely to be very low, which may be intuitively misleading about the actual pertinence of the explanatory variables. For example, in Pélissier *et al.* (2003), we found that a soil gradient coded in nine classes, though highly significant (randomization test: $P < 0.001$; see below), explained less than 5% of the Simpson diversity of a table of species occurrences of 381 individuals and 113 species.

An appropriate statistic to test the null hypothesis of no effect of the explanatory variables is thus the ANOVA-like pseudo-*F* ratio (Legendre & Anderson 1999), which includes the degrees of freedom in the numerator and denominator of the R^2 ratio. We call it '*pseudo*' because the theoretical distribution function of this statistic is unknown and probably not a Fisher-Snedecor distribution, as $\mathbf{Y}$ does not conform to a multinormal distribution function. Non-parametric tests of statistical significance such as those based on randomization procedures (Anderson 2001; McArdle & Anderson 2001) are therefore required. Indeed, an empirical distribution of the pseudo-*F* ratio can be simply obtained by permutations between the rows of $\mathbf{Y}$, which have uniform weights of $1/n$, while species weights are kept unchanged.

Relationships with alpha-, beta- and gamma-diversity

Our definition of total species diversity, *TSS* in eqn 1, obviously conforms to Whittaker's (1972) concept of gamma-diversity as a measure of species diversity in a pooled set of samples, i.e. from '… samples combined from several communities, or lists of species for geographical units, or nonareal samples […] drawing species from a number of communities'. Whittaker also postulated that: '… the extent of change in species composition of communities […] along environmental gradients is beta diversity or between-habitat diversity'. However, since then, beta-diversity is usually viewed as a measure of the variation in species composition between discrete samples (Magurran 2004), such as, study sites or habitat types (e.g. soil classes). Our multivariate linear model provides in this case a direct generalization of Lande's (1996) partition within the framework of (multi)factorial multivariate analysis of variance (see first subsection below). However, while the environmental distance between groups of observations is arbitrary and constant in a factorial experimental design, our model also provides a means for the direct quantification of gradient-induced beta-diversity when the sampling points are placed with respect to a continuous environmental variable (second subsection).

DISCRETE HABITAT TYPES

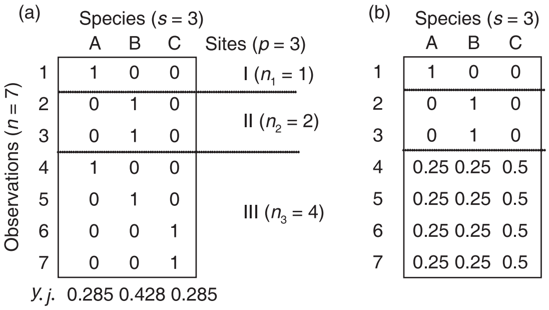
Returning to eqn 2a with a hypothetical example similar to the one of Fig. 1(b): $\mathbf{Y}$ is a $n \times s$ table of species occurrences and $\mathbf{X}$ a $n \times (p-1)$ matrix of dummy variables coding for an explanatory categorical descriptor with p habitat types, environmental classes or sampling sites (see Legendre & Legendre 1998, p. 46). Couteron & Pélissier (2004) showed that such a model enters within the framework of multivariate analysis of variance, *sensu* Anderson (2001), i.e. a generalization of the univariate ANOVA obtained by adding up the sum of squares across all dependent variables. Indeed, we can re-formulate the table of species occurrences in order to take explicitly into account the partition of the n observations into p sites as $\mathbf{Y}$, whose elements are noted y_{ijk}, with $1 \le i \le n$, $1 \le j \le s$ and $1 \le k \le p$. The total number of observations is $n = \sum_k n_k$, where n_k is the number of observations in site k.

In doing so, the approximated values of $\mathbf{Y}$ by $\mathbf{X}$, noted $\hat{y}_{ijk}$, are the mean relative frequencies of the species within each site, namely $y_{\cdot jk} = \sum_{i \in k} y_{ijk}/n_k$, so that the approximated occurrence table $\hat{\mathbf{Y}}$, whose rows are all the same in a given class k (Fig. 2), is unbiased:

$$y_{\cdot j\cdot} = \bar{\hat{y}}_{ijk} = \sum_i \hat{y}_{ijk}/n = \sum_k n_k y_{\cdot jk}/n$$

Therefore, expressing *MSS* and *RSS* as the among- and within-sites sum of squares gives:

$$MSS = \sum_j w_j \sum_i (\hat{y}_{ijk} - y_{\cdot j\cdot})^2 = \sum_j w_j \sum_k n_k (y_{\cdot jk} - y_{\cdot j\cdot})^2 \qquad \text{eqn 3a}$$



(a)

Observations ($n = 7$)	Species ($s = 3$) A	B	C	Sites ($p = 3$)
1	1	0	0	I ($n_1 = 1$)
2	0	1	0	II ($n_2 = 2$)
3	0	1	0	
4	1	0	0	III ($n_3 = 4$)
5	0	1	0	
6	0	0	1	
7	0	0	1	
$y._j$	0.285	0.428	0.285	

(b)

	Species ($s = 3$) A	B	C
1	1	0	0
2	0	1	0
3	0	1	0
4	0.25	0.25	0.5
5	0.25	0.25	0.5
6	0.25	0.25	0.5
7	0.25	0.25	0.5

Fig. 2 A hypothetical example of diversity partitioning with respect to discrete habitat types using standard (M)ANOVA routines. (a) Initial table of species occurrences, **Y**. (b) Approximated table by a set of dummy variables coding for discrete habitat types (or sites), $\hat{\mathbf{Y}}$.

$$RSS = \sum_j w_j \sum_i (y_{ijk} - \hat{y}_{ijk})^2$$
$$= \sum_j w_j \sum_k \sum_{i \in k} (y_{ijk} - y_{\cdot jk})^2 \qquad \text{eqn 3b}$$

Dividing the above equations by n renders them equivalent to those defining beta- and alpha-diversity, respectively, in the additive partition of Lande (1996) or Couteron & Pélissier (2004).

Contrary to the assertion of Crist *et al.* (2003), it is here demonstrated that any statistical package dealing with ANOVA can provide additive apportionment of species diversity within beta and alpha components, namely *MSS*/*n* and *RSS*/*n*. In fact, the results shown in Table 1 were obtained through function *aov* ( ) (Appendix S1 in Supplementary Material) of the *R* statistical package (R Development Core Team 2004). Options for two-way ANOVA, which are available with the same functions, can address more sophisticated schemes of diversity partitioning as presented by Couteron & Pélissier (2004). The approach of permutation tests based on the pseudo-*F* ratio remains useful in this context. The guidelines provided by Anderson & Ter Braak (2003) provide a sound basis although technical investigations on power and accuracy of these tests are still needed in the case of multiway and/or nested ANOVA. The influence of species weighting on power and accuracy of these tests is an open question, which should also be addressed.

CONTINUOUS ENVIRONMENTAL GRADIENT

Let us now consider a $n \times s$ table of species occurrences and a continuous variable **X** corresponding to a quantitative measure of an ecological characteristic (e.g. soil pH) recorded for each site or relevé. The amount of variation in **Y** accounted for by the variation of **X** is thus quantified, in any diversity metric defined via w_j, by:

$$MSS = \sum_j w_j \cdot (\sum_i (x_i - \bar{x}) \cdot (y_{ij} - y_{\cdot j}))^2 / \sum_i (x_i - \bar{x})^2 \qquad \text{eqn 4}$$

It follows that *MSS*/*n* represents the part of total species diversity explained by the gradient, i.e. an objective measure of the gradient-induced beta-diversity.

Imagine, for instance, that soil pH was 4.6, 5.3 and 5.8 for the three sites of our hypothetical example, respectively. Any statistical package dealing with linear models can provide the results given in Table 2 and obtained using the *aov* ( ) wrapper function to *lm* ( ) (Appendix S1) of the *R* statistical package (R Development Core Team 2004).

By extension, multivariate analysis of covariance provides a means to adjust for the effects of a continuous covariate in an ANOVA design (Sokal & Rohlf 1995).

Relationships with distance/dissimilarity matrices

Since Whittaker, beta-diversity is often quantified by distance (or dissimilarity) matrices derived from various similarity coefficients (reviewed by Legendre & Legendre 1998, p. 253). Unfortunately, the most frequently used similarity indices (e.g. Jaccard, Sorensen or Steinhaus) have no direct connection with the usual

Table 1 Diversity partitioning with respect to discrete habitat types using standard (M)ANOVA routines and the hypothetical example given in Fig. 2(a)

	Total diversity (*TSS*/*n*)	Total diversity (*MSS*/*n*)	R^2 (*MSS*/*TSS*)	Pseudo-*F* $\dfrac{MSS/(p-1)}{(TSS - MSS)/(n - p)}$
Richness – 1	2	0.875	0.4375	1.56
Shannon	1.08	0.482	0.4464	1.61
Simpson	0.653	0.296	0.4533	1.66

Table 2 Diversity partitioning with respect to a continuous environmental gradient using standard (M)ANOVA routines and the hypothetical example given in Fig. 2(a) with pH values of 4.6, 5.3 and 5.8 assigned to sites I, II and III, respectively

	Total diversity (*TSS*/*n*)	Total diversity (*MSS*/*n*)	R^2 (*MSS*/*TSS*)	Pseudo-*F* $\dfrac{MSS/(p-1)}{(TSS - MSS)/(n - p)}$
Richness – 1	2	0.29	0.145	0.847
Shannon	1.08	0.145	0.134	0.778
Simpson	0.653	0.0829	0.127	0.727

diversity indices, which means that many ecological studies measured alpha and beta diversity in distinct 'units', a somewhat unsatisfying situation. Moreover, as recently pointed out by Legendre *et al.* (2005), some confusion has risen in the literature concerning the possible relationship between the measure of beta diversity and the variance of an abundance data table. We will first examine this in more detail, drawing upon connection with multivariate ordination techniques. We will then show how our model can help clarify the relationship between dissimilarity and beta diversity, and thus provide a basis for more consistent spatially explicit apportionments of species diversity.

FROM OCCURRENCES TO ABUNDANCES

Let us refer to an arbitrary $p \times s$ abundance matrix, **A**, with sites as rows ($1 \le k \le p$) and species as columns ($1 \le j \le s$). As shown in Fig. 1(b), such a table is closely related to **Y**, our table of species occurrences. Similarly, an '*ecologically meaningful transformation*' of abundances into '*compositional data*' (Legendre & Gallagher 2001) as $c_{kj} = a_{kj}/n_k$, provides a 'shrunken' version of $\hat{\mathbf{Y}}$, the approximation of **Y** by a set of dummy variables coding for sites, without any loss of information, since we have for each k: $\hat{y}_{ijk} = \sum_{i \in k} y_{ijk}/n_k = a_{kj}/n_k$ (see previous section and Fig. 3a).

However, the classical sum of squares, i.e. the sum of the squared deviations from the mean, of the transformed **C** matrix, is not equivalent to *MSS* computed from the occurrences. It indeed appears from eqn 3a, that *MSS* can be viewed as a weighted sum of the squared differences between within-sites and overall relative species frequencies, that may be expressed as:

$$MSS = \sum_j w_j \sum_k n_k(a_{kj}/n_k - y_{\cdot j \cdot})^2 \qquad \text{eqn 5a}$$

with $y_{\cdot j \cdot} = \sum_k n_k y_{\cdot jk}/n = \sum_k n_k a_{kj}/n_k n = \sum_k a_{kj}/n$

In the above equation, it is as if the values of the abundance table, **A**, have been re-scaled thanks to a division by n_k (in matrix **C**), while the rows (sites) have been provided a weight of n_k, and the columns (species) a weight of w_j. By dividing *MSS* by n, one can thus recognize an expression of the *total inertia* (or total variance, i.e. the sum of all eigenvalues) of correspondence analysis (CA) when $w_j = 1/y_{\cdot j \cdot}$, and of a form of redundancy analysis (RDA) called non-symmetric correspondence analysis (NSCA) when $w_j = 1$ (Gimaret-Carpentier *et al.* 1998; Pélissier *et al.* 2003). Taking $w_j = \log(1/y_{\cdot j \cdot})/(1 - y_{\cdot j \cdot})$, could also lead to a form of *column weighted* correspondence analysis whose inertia is consistent with Shannon diversity (see Pélissier *et al.* 2003). Other alternatives for re-scaling and row weighting consistent with well-known and useful ordination methods are possible, although they are in this case incompatible with usual diversity indices (Couteron & Ollier 2005).

(a)

Sites (p = 3)	A	B	C	n_k
I	1	0	0	1
II	0	1	0	2
III	0.25	0.25	0.5	4

Species (s = 3)

(b)

	A	B	C	$n_k n_{k'}$
$D^2_{(\mathrm{I,\,II})}$	1	1	0	2
$D^2_{(\mathrm{I,\,III})}$	0.5625	0.0625	0.25	4
$D^2_{(\mathrm{II,\,III})}$	0.0625	0.5625	0.25	8

Species (s = 3)

Fig. 3 (a) The 'shrunken' matrix **C**, containing the within-sites profiles of relative species frequencies, for the hypothetical example given in Fig. 2. (b) Squared Euclidean distances between sites or dissimilarities (provided here $w_j = 1$ for all species, i.e. using Simpson metric).

Various *R* packages, available from the CRAN repository (see Appendix S1), offer functions to perform multivariate ordinations that retrieve the results of Table 1, such as *corresp* ( ) of package *MASS*, *cca* ( ) of package *vegan*, *dudi.coa* ( ) and *dudi.nsc* ( ) of package *ade4*.

FROM ABUNDANCES TO DISSIMILARITIES

In addition, we can express *MSS* as the mean of the squared Euclidean distances among the *n* observations (Legendre & Anderson 1999). This means that averaging squared departures around a mean value is equivalent to averaging squared differences between individual observations (see Anderson 2001). In so doing, *MSS* of eqn 3a can be rewritten as:

$$\begin{aligned} MSS &= \sum_j w_j \sum_i \sum_{i'} (\hat{y}_{ij} - \hat{y}_{i'j})^2/2n \\ &= \sum_j w_j \sum_k \sum_{k'} \sum_{i \in k} \sum_{i' \in k'} (\hat{y}_{ijk} - \hat{y}_{i'jk'})^2/2n \end{aligned} \qquad \text{eqn 5b}$$

For reverting to abundances, we remember that $\hat{y}_{ijk} = a_{kj}/n_k$ for all observations belonging to a given site k, so that:

$$MSS = \sum_j w_j \sum_k \sum_{k'} n_k n_{k'}(a_{kj}/n_k - a_{k'j}/n_{k'})^2/2n \qquad \text{eqn 5c}$$

It is apparent from eqn 5c that *MSS* is a weighted average of the squared '*distance between species profiles*' (Legendre & Gallagher 2001) of sites k and k', i.e. $(a_{kj}/n_k - a_{k'j}/n_{k'})^2$, and that sites are weighted according to the number of occurrences they harbour while species weighting, w_j, defines the diversity metric. Furthermore, one can build a measure of dissimilarity between sites k and k', $D^2_{(k,k')} = \sum_j w_j(a_{kj}/n_k - a_{k'j}/n_{k'})^2$, which is consistent with any of the diversity metrics defined by w_j, as:

$$\begin{aligned} MSS &= \sum_k \sum_{k'} n_k n_{k'} D^2_{(k,k')}/2n \\ &= \sum_{k=1}^{p-1} \sum_{k'=k+1}^{p} n_k n_{k'} D^2_{(k,k')}/n \end{aligned} \qquad \text{eqn 5d}$$

Dissimilarities given in Fig. 3(b) were obtained thanks to the standard *dist* ( ) function (Appendix S1) in *R* statistical package (R Development Core Team 2004).

Couteron & Pélissier (2004) and Couteron & Ollier (2005) demonstrated that various subsequent spatially explicit apportionments of species diversity are derived from eqn 5c, on the basis of ecological and/or geographical distance classes among sites.

Conclusion and perspectives

What are the relative effects of the biological, environmental and anthropogenic factors, and of their potential interactions on species diversity? Are these effects stable across scales, from landscape to region, between regions and across ecosystems?

We have presented here a simple multivariate linear model, which enables us to address these questions by partitioning the most common diversity indices according to environmental explanatory variables on the basis of standard, well-mastered methods of variance and covariance decomposition. Thanks to an unusual form of presentation of the taxonomic data, the table of species occurrences, which considers individual organisms as the elementary statistical unit, this approach extends and generalizes the principles of additive partitioning (Lande 1996) and hierarchical analysis (Wagner *et al.* 2000; Crist *et al.* 2003) of species diversity. An additional practical advantage is that standard functions of (multivariate) analysis of variance, such as the *aov* ( ) function in *R*, can directly be used to perform the computations. However, given that a table of species occurrences may be very large, and with a high proportion of zero entries (i.e. a sparse matrix, Duff *et al.* 1986), optimized dedicated *R* routines have been made freely available at http://pelissier.free.fr/Diversity.html. The code to perform the worked examples provided in this paper with both standard *R* functions and our *diversity* routines is given in Appendix S1.

Conforming to a standard analytical framework provides an interesting perspective on a variety of analyses of the components of species diversity, which preserves consistency with the common richness, Shannon and Simpson diversity indices. We showed for instance, that ordination in reduced space (a form of variance apportioning) of the fitted and residual tables of our model had direct links with correspondence analysis and with some of its one- or two-table variants, such as canonical correspondence analysis (Pélissier *et al.* 2003). Moreover, spatially explicit diversity partitioning can be related to variography, a form of variance decomposition in relation to distance (Couteron & Pélissier 2004), which further extends towards the analysis of spatial patterns displayed by multivariate ordination results (multiscale ordination, Wagner 2003; Couteron & Ollier 2005).

Such a methodological integration provides a means to conduct various complementary analyses in the same diversity metric, and in particular to measure alpha and beta diversity components in the same unit. This facilitates assessing the relative effects of different types of processes affecting species diversity patterns, as well as investigating their stability across scales.

However, in terms of methodology, more remains to be explored. For instance, a well-known drawback of parametric, as well as simple randomization procedures to test for statistical significance of the ANOVA-like *F*-ratios, is the underlying assumption of independence between the observations. Data collected via taxonomic relevés are likely to violate this assumption because of spatial autocorrelation of species' distributions, a fact that could yield undue significance of effects of certain explanatory variables (Legendre & Fortin 1989; Legendre 1993). Permutation strategies accounting for the spatial structure of multifactorial experimental designs are available (Anderson & Ter Braak 2003), but they still do not meet the hypothesis of independence between the individual observations at the lowest sampling strata. Techniques borrowed from geostatistics that incorporate spatial dependence as an additional term within a standard linear model, seem promising (Lichstein *et al.* 2002; Wall 2004), as the term of spatial dependence is able to represent processes endogenous to vegetation dynamics (dispersal, demography, etc.), i.e. with no strict environmental determinism (Keitt *et al.* 2002).

Up to this point, we have only discussed explanatory models that seek to account for observed variations in species diversity. However, one could also want to predict species' occurrences and diversity at unsampled locations, a more demanding objective. In this case, the variables to be predicted are the Y_i columns of the table of the species occurrences that are binomial variables. A natural refinement of our model that constrains the predictions to be probabilities of species occurrences ranging between 0 and 1, is a multivariate logistic model (Hosmer & Lemeshow 2000), i.e. a generalized multivariate linear model with a logit link function, classically used to predict presence-absence data (e.g. Dupré & Ehrlén 2002; Guisan *et al.* 2002; Kolb & Diekmann 2004; Guisan & Thuiller 2005). Spatial dependence can also be introduced under the form of geostatistical or of conditional autoregressive models (CAR) that can suit the prediction of binary variables (Anselin 2002).

While ongoing biodiversity census and development of information technologies will ease the constitution of large and relevant data sets, modelling diversity determinants and variations will demand appropriate statistical standards for data analysis and parameters estimations. We hope our contribution will stimulate further developments in this way.

Acknowledgements

This research has been carried out in the framework of the OSDA (Organization Spatiale de la Diversité des Arbres) project supported by the French Ministère de l'Ecologie et du Développement Durable. UMR AMAP (Botany and Bioinformatics applied to Plant Architecture) is a joint research unit, which associates CIRAD (UMR51), CNRS (UMR5120), INRA (UMR931), IRD 2M123, and Montpellier 2 University (UM27). The French Institute of Pondicherry (IFP) is a research centre of the French Ministry of Foreign Affairs. We are very grateful to D. Chessel from the University of Lyon for important insights into the relationships between diversity indices and multivariate analyses, and to F. Houllier from INRA for pointing out to us the connection with predictive species distribution models.

References

Anderson, M.J. (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, **26**, 32–46.

Anderson, M.J. & Ter Braak, C.J.F. (2003) Permutation tests for multi-factorial analysis of variance. *Journal of Statistical Computation and Simulation*, **73**, 85–113.

Anselin, L. (2002) Under the hood. Issues in the specification and interpretation of spatial regression models. *Agricultural Economics*, **27**, 247–267.

Couteron, P. & Ollier, S. (2005) A generalized, variogram-based framework for multiscale ordination. *Ecology*, **86**, 828–834.

Couteron, P. & Pélissier, R. (2004) Additive partitioning of species diversity: towards more sophisticated models and analyses. *Oikos*, **107**, 215–221.

Crist, T.O., Veech, J.A., Gering, J.C. & Summerville, K.S. (2003) Partitioning species diversity across landscapes and regions: a hierarchical analysis of alpha-, beta-, and gamma-diversity. *American Naturalist*, **162**, 734–743.

Duff, I.S., Erisman, A.M. & Reid, J.K. (1986) *Direct Methods for Sparse Matrices*. Clarendon Press, Oxford.

Dupré, C. & Ehrlén, J. (2002) Habitat configuration, species traits and plant distribution. *Journal of Ecology*, **90**, 796–805.

Gaston, K.J. & Blackburn, T.M. (2000) *Pattern and Process in Macroecology*. Blackwell Science, Oxford.

Gimaret-Carpentier, C., Chessel, D. & Pascal, J.-P. (1998) Non-symmetric correspondence analysis: an alternative for species occurrences data. *Plant Ecology*, **138**, 97–112.

Guisan, A., Edwards, T.C.J. & Hastie, T. (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling*, **157**, 89–100.

Guisan, A. & Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, **8**, 993–1009.

Hosmer, D.W. & Lemeshow, S. (2000) *Applied Logistic Regression*. John Wiley & Sons, New York.

Keitt, T.H., Bjornstad, O.N., Dixon, P.M. & Citron-Pousty, S. (2002) Accounting for spatial pattern when modeling organism–environment interactions. *Ecography*, **25**, 616–625.

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

Kolb, A. & Diekmann, M. (2004) Effects of environment, habitat configuration and forest continuity on the distribution of forest plant species. *Journal of Vegetation Science*, **15**, 199–208.

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

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

Legendre, P. & Anderson, M.J. (1999) Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs*, **69**, 1–24.

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

Legendre, P. & Fortin, M. (1989) Spatial pattern and ecological analysis. *Vegetatio*, **80**, 107–138.

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

Legendre, P. & Legendre, L. (1998) *Numerical Ecology*. Elsevier, Amsterdam, the Netherlands.

Lichstein, J.W., Simons, T.R., Shriner, S.A. & Franzreb, K.E. (2002) Spatial autocorrelation and autoregressive models in ecology. *Ecological Monographs*, **72**, 445–463.

Magurran, A.E. (2004) *Measuring Biological Diversity*. Blackwell Science, Malden, MA.

McArdle, B.H. & Anderson, M.J. (2001) Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*, **82**, 290–297.

Pélissier, 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.

R Development Core Team (2004) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org.

Sokal, R.R. & Rohlf, F.J. (1995) *Biometry: the Principles and Practice of Statistics in Biological Research*. Freeman, San Francisco, CA.

Veech, J.A., Summerville, K.S., Crist, T.O. & Gering, J.C. (2002) The additive partitioning of species diversity: recent revival of an old idea. *Oikos*, **99**, 3–9.

Wackernagel, H. (1998) *Multivariate Geostatistics*. Springer-Verlag, Berlin, Germany.

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

Wagner, H.H., Wildi, O. & Ewald, K.C. (2000) Additive partitioning of plant species diversity in an agricultural mosaic landscape. *Landscape Ecology*, **15**, 219–227.

Wall, M.M. (2004) A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference*, **121**, 311–324.

Whittaker, R.H. (1960) Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs*, **30**, 279–338.

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

Yoccoz, N., Nichols, J.D. & Boulinier, T. (2001) Monitoring of biological diversity in space and time. *Trends in Ecology and Evolution*, **16**, 446–453.

Received 24 July 2006 revision accepted 27 October 2006
Handling Editor: Rob Brooker

Supplementary material

The following supplementary material is available for this article:

Appendix S1 R code to perform the worked examples.

This material is available as part of the online article from: http://www.blackwell-synergy.com/doi/abs/10.1111/j.1365-2745.2007.01211.x

Please note: Blackwell Publishing is not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Appendix

R code to perform the worked examples with the standard functions of the *base* library or other libraries (*stats, ade4, vegan*) available at http://cran.r-project.org/, and dedicated *diversity* routines available at http://pelissier.free.fr/Diversity.html.

`spot` is the table of species occurrences, `hat` the vector of discrete habitat types and `ph` the vector of soil pH variable.

```
> spot<-cbind(c(1,0,0,1,0,0,0),c(0,1,1,0,1,0,0),c(0,0,0,0,0,1,1))
> hat<-as.factor(c(1,2,2,3,3,3,3))
> ph<-c(4.6,5.3,5.3,5.8,5.8,5.8,5.8)
```

`n` is the total number of observations.

```
> n<-sum(spot)
```

`wrich` and `wshan` are vectors of species weights for the richness and Shannon metrics, respectively (for Simpson metric the weight is 1 for all species).

```
> wrich<-1/apply(spot,2,mean)
> wshan<-log(1/apply(spot,2,mean))/(1-apply(spot,2,mean))
```

1. Partitioning species diversity according to discrete habitat types (Tab. 1):

1a. the standard `aov{base}` function performs (M)ANOVA, while the `summary.div{diversity}` function converts sum of squares into diversity measures:

```
> library(diversity)
> summary.div(aov(spot~hat))
```

1b. the `anodiv{diversity}` function performs MANOVA from a specific `odf` object more efficient for large data sets. Moreover, `summary{diversity}` performs randomisation tests of statistical significance:

```
> ovec<-odf(spot)
> summary(anodiv(ovec$sp~hat),nrepet=100)
```

2. Partitioning species diversity according to pH gradient (Tab. 2) with the standard `aov{base}` function and `summary.div{diversity}` (randomisation tests for continuous covariates are not yet implemented in package `diversity`):

```
> summary.div(aov(spot~ph))
```

3. Computing *MSS* from inertia of an abundance data table (Tab. 1):

3a. `corresp{MASS}` function performs Correspondence Analysis (CA) providing *MSS* in the richness metric:

```
> A<-apply(spot,2,function(x) tapply(x,hat,sum))
> library(MASS)
> sum(corresp(A,nf=2)$cor^2)
```

3b. `dudi.coa{ade4}` and `dudi.nsc{ade4}` perform CA and Non-Symmetric CA providing *MSS* in both richness and Simpson metrics, respectively:

```
> library(ade4)
> sum(dudi.coa(as.data.frame(A),scannf=F)$eig)
> sum(dudi.nsc(as.data.frame(A),scannf=F)$eig/ncol(A))
```

3c. `ca.richness{diversity}`, `nsca.simpson{diversity}` and `cwca.shannon{diversity}` are wrapper functions to `as.dudi{ade4}`, which perform *MSS* in the richness, Shannon and Simpson metrics, respectively:

```
> rich<-ca.richness(A) #Select the number of axes:

> 2

> summary(rich)


> shan<-cwca.shannon(A) #Select the number of axes:

> 2

> summary(shan)


> simp<-nsca.simpson(A) #Select the number of axes:

> 2

> summary(simp)
```

4. Computing *MSS* from dissimilarity matrices (Fig. 3):

4a. `dist{base}` function computes Euclidean distances:

```
> ni<-apply(A,1,sum)

> w<-c(ni[1]*ni[2],ni[1]*ni[3],ni[2]*ni[3])

> B<-apply(A,2,function(x) x/ni)

> D2<-apply(B,2,dist)^2

> rich.ssq<-sum(apply(t(apply(D2,1,function(x) x*wrich)),2,function(x) x*w))/n^2

> shan.ssq<-sum(apply(t(apply(D2,1,function(x) x*wshan)),2,function(x) x*w))/n^2

> simp.ssq<-sum(D2*w)/n^2

> cbind(rich.ssq,shan.ssq,simp.ssq)
```

4b. `dissim{diversity}` function computes dissimilarities:

```
> d<-dissim(A)

> D2<-d$delta

> apply(d$dissim,2,function(x) sum(x*d$weight))
```