Ecology, 84(1), 2003, pp. 242–251

CONSISTENCY BETWEEN ORDINATION TECHNIQUES AND DIVERSITY MEASUREMENTS: TWO STRATEGIES FOR SPECIES OCCURRENCE DATA

RAPHAËL PÉLISSIER,[1,4] PIERRE COUTERON,[2] STÉPHANE DRAY,[3] AND DANIEL SABATIER[1]

[1]*IRD, UMR botAnique et bioinforMatique de l'Architecture des Plantes, TA40/PS2, 34398 Montpellier Cedex 05, France*
[2]*ENGREF, UMR botAnique et bioinforMatique de l'Architecture des Plantes, TA40/PS2, 34398 Montpellier Cedex 05, France*
[3]*Université Lyon 1, UMR Biométrie et Biologie Evolutive, 69622 Villeurbanne Cedex, France*

Abstract. Both the ordination of taxonomic tables and the measurements of species diversity aim to capture the prominent features of the species composition of a community. However, interrelations between ordination techniques and diversity measurements are seldom explicated and are mainly ignored by many field ecologists. This paper starts from the notion of the species occurrence table, which provides a unifying formulation for different kinds of taxonomic data. Here it is demonstrated that alternative species weightings can be used to equate the total inertia of a centered-by-species occurrence table with common diversity indices, such as species richness, Simpson diversity, or Shannon information. Such an equation defines two main ordination strategies related to two different but consistent measures of species diversity. The first places emphasis on scarce species and is based on Correspondence Analysis and species richness (CA-richness strategy). The second, in which abundant species are prominent, relies on Non-Symmetric Correspondence Analysis and Simpson diversity (NSCA-Simpson strategy). Both strategies are suitable for measuring α and β diversity by analyzing the centered-by-species occurrence table with respect to external environmental or instrumental variables.

In this paper, these two strategies are applied to ecological data obtained in a Neotropical rainforest plot. The results are then discussed with respect to the intrinsic characteristics of the community under analysis, and also to the broad classes of floro-faunistic data used in ecology (i.e., data gathered from museum or herbarium collections, exhaustive inventories in a reference plot, or enumeration through species-by-relevés tables). The approach encompasses several well-known techniques such as Correspondence Analysis, Non-Symmetric Correspondence Analysis, Canonical Correspondence Analysis, and Redundancy Analysis, and provides greater insight into interrelations between ordination methods and diversity studies.

Key words: *α and β diversity; inertia decomposition; multivariate analysis; Simpson diversity; species–environment relationships; species occurrence table; species richness; species weight.*

INTRODUCTION

The measurement of species diversity is a central topic in community ecology (Ricklefs 1990, Krebs 1994, Begon et al. 1996). The simplest measure of diversity is species richness, which corresponds to the number of species present in a community. But numerous other indices, based on the idea that species frequency distribution is more informative than simple species richness, can be found in ecological literature (e.g., Magurran 1988). Of these, the famous nonparametric Shannon (1948) information, H, Simpson (1949) concentration, λ, and Simpson diversity, $D = 1 - \lambda$ (Greenberg 1956) are the most commonly employed. Pielou (1969) has shown that D is an unbiased sample estimator of population diversity that is insensitive to very rare species (which are generally poorly sampled) but sensitive to changes in the abundance of the few prevalent species, so that Hill (1973) recommended the use of $1/\lambda$ instead. Following Patil and Taillie (1982), Lande (1996) noted however that Simpson $D = 1 - \lambda$ can be expressed as the total variance in species identity within a community.

Whittaker (1972) introduced the important concept of diversity partitioning into within (α) and between (β) components, which prefigured an interrelation between ordination techniques and diversity studies (Gauch and Whittaker 1972, Gauch 1973). However, for most field ecologists, this link remained somewhat abstract because it was not clearly related to the commonly used indices of species diversity. Later, Ter Braak (1983) emphasized that Principal Component Analysis (PCA) based on species profiles can be interpreted in terms of α and β diversity related to Simpson D. More recently, Gimaret-Carpentier et al. (1998*a*) pointed out that total inertia computed by Non-Symmetric Correspondence Analysis (NSCA; Lauro and D'Ambra 1984) corresponds exactly to Simpson diversity. As far as we know, however, no paper has

Manuscript received 27 September 2001; revised 5 April 2002; accepted 15 May 2002; final version received 21 June 2002. Corresponding Editor: D. W. Roberts.
[4] E-mail: Raphael.Pelissier@mpl.ird.fr

ever been published in the ecological literature to really unify the notions of ordination and diversity measures.

This paper starts from the notion of the species occurrence table, which is a unifying formulation for different kinds of taxonomic data. It will be shown that the choice of appropriate species weighting modifies the metric of the eigenanalysis of a centered-by-species occurrence table and enables its total inertia to be equated with the common diversity indices. Such an equation will be extended to α and β diversity by considering a constraint analysis (The CoCoAn [Constrained Correspondence Analysis] R Package; Dray 2001, available online)[5] of the occurrence table with respect to environmental or instrumental (sensu, Rao 1964) variables. Our aim is to underline how the analysis of species–environment relationships may rely on two main alternative strategies that encompass simultaneously an ordination technique and a consistent measurement of taxonomic diversity. Both strategies will be applied to ecological data obtained in a Neotropical rainforest plot, before being discussed with respect to the broad classes of floro-faunistic data used in ecology.

Measuring Total Diversity from a Table of Species Occurrences

A table of species occurrences is a very simple, though unusual, form of presentation of taxonomic data, where each individual recorded is allocated to a taxonomic category (usually a particular species). In terms of data analysis, it corresponds to a complete binary table **T** with n rows (individuals or occurrences) and p columns (species), such that

$$\mathbf{T}_{(n\times p)} = [t_{ij}] = \begin{cases} 1 & \text{if the } i\text{th occurrence belongs} \\ & \quad \text{to the } j\text{th species} \\ 0 & \text{otherwise.} \end{cases}$$

Table **T** differs from the usual site-by-species ecological table since it is irrespective of sampling units: each row represents an individual organism identified to species and observed either at a single site or throughout a collection of sites. Table **T** has some analogy to the ''inflated matrix'' (Legendre and Legendre 1998: p. 595) used in Canonical Correspondence Analysis (CCA; Ter Braak 1986) to get a weighted regression. Such a species occurrence table may originate from data gathered from museum or herbarium collections (e.g., Gimaret-Carpentier et al. 1998*a*), exhaustive inventories in a reference plot (e.g., Pélissier et al. 2002), or species enumeration through several relevés.

Column means of **T** are simply the species relative frequencies, noted $f_j = t_{+j}/t_{++}$, where t_{+j} is the sum of values in column j (i.e., the number of occurrences belonging to species j), and t_{++} is the sum of all values in the whole table (i.e., n, the total number of occurrences). Let us consider **Tc** of size $(n \times p)$, the centered-by-columns (species) table derived from **T**. It contains either $t_{ij} - f_j = 1 - f_j$ when the occurrence belongs to species j (i.e., $t_{ij} = 1$) or $t_{ij} - f_j = -f_j$ when the occurrence does not belong to species j (i.e., $t_{ij} = 0$). Column means of **Tc** are zero and column variances are $f_j(1 - f_j)$.

Let us now consider a $(n \times n)$ diagonal matrix of row weights, noted $\mathbf{D}_n$, and a $(p \times p)$ diagonal matrix of species weights, noted $\mathbf{\Delta}_p$. Matrix $\mathbf{D}_n$ is defined such that

$$\mathbf{D}_n = [d_{n_{ij}}] = \begin{cases} f_i = t_{i+}/t_{++} & i = j \\ 0 & i \neq j. \end{cases}$$

For the table of species occurrences **T** defined above, $t_{++} = n$ and $\forall i,\ t_{i+} = 1$, hence $\mathbf{D}_n$ contains on its diagonal the natural weights of the occurrences (i.e., $f_i = 1/n$). Our purpose is first to show that varying the species weights contained in $\mathbf{\Delta}_p$ changes the metric of the eigenanalysis of table **Tc** and gives it some interesting properties with respect to diversity measurements.

For the sake of clarity, the statistical triplet notation of Escoufier (1987) will be adopted to specify the type of analysis under consideration (see also Dolédec et al. 1996, 2000, for an introduction to triplet notation in ecology). For instance, the eigenanalysis of the statistical triplet (**Tc**, $\mathbf{\Delta}_p$, $\mathbf{D}_n$) indicates that we perform a generalized singular value decomposition (GSVD; Greenacre 1984) of table **Tc**, using column weights given by $\mathbf{\Delta}_p$ and row weights given by $\mathbf{D}_n$ (see Appendix A).

Let us consider first the classical χ^2 metric of the Correspondence Analysis (CA; Hill 1974) that defines a diagonal matrix $\mathbf{\Delta}_p = \mathbf{D}_p^{-1}$ containing species weights corresponding to

$$\mathbf{D}_p^{-1} = [d_{p_{ij}}] = \begin{cases} 1/f_j & i = j \\ 0 & i \neq j. \end{cases}$$

The CA of **T** can be defined from the statistical triplet (**Tc**, $\mathbf{D}_p^{-1}$, $\mathbf{D}_n$), whose eigenanalysis has total inertia corresponding to the weighted sum of the column variances of **Tc**:

$$I_{\mathrm{T}} = \sum_{j=1}^{p} \frac{1}{f_j} \times f_j(1 - f_j) = \sum_{j=1}^{p} (1 - f_j) = p - 1$$

where p corresponds to the total number of species (columns of **Tc**). Species richness expressed as $p - 1$ is obviously a measure of diversity (Hill 1973, Patil and Taillie 1982) since a community containing a single species has a diversity of zero. Note also that the lower the value of f_j, the greater the contribution by a given species (Fig. 1).

Consider now a diagonal identity matrix $\mathbf{\Delta}_p = \mathbf{I}_p$ that contains uniform species weights:

[5] URL: ⟨http://lib.stat.cmu.edu/R/CRAN/src/contrib/PACKAGES.html#CoCoAn⟩

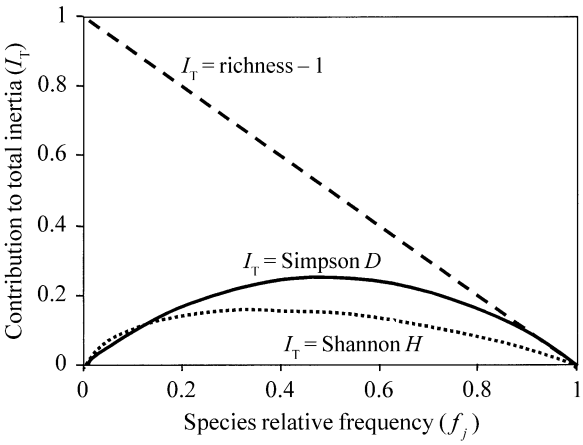

FIG. 1. Species contributions to the total inertia (I_T) of the centered-by-species occurrence table (**Tc**) in relation to the relative frequency of species (f_j) when species weights are: $1/f_j$ (I_T = species richness -1); $\log[1/f_j]/(1 - f_j)$ (I_T = Shannon information H); or 1 (I_T = Simpson diversity D).

$$\mathbf{I}_p = [i_{p_{ij}}] = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases}$$

The Non-Symmetric Correspondence Analysis (NSCA) of **T** can be defined from the statistical triplet (**Tc**, $\mathbf{I}_p$, $\mathbf{D}_n$), whose eigenanalysis uses the Euclidian metric, and corresponds simply to the centered Principal Component Analysis (PCA) of **T**. The total inertia of this analysis computed as the weighted sum of the column variances of **Tc** is

$$I_T = \sum_{j=1}^{p} f_j(1 - f_j) = 1 - \sum_{j=1}^{p} f_j^2$$

which corresponds exactly to Simpson diversity. The contribution of a given species increases here with its relative frequency on condition that f_j does not reach 0.5 (Fig. 1).

It should also be noted that $\mathbf{\Delta}_p$ containing species weights of $\log[1/f_j]/(1 - f_j)$ leads to an intermediate situation between species richness and Simpson index (Fig. 1; see also Gimaret-Carpentier et al. 1998*b*), where the total inertia of the eigenanalysis of (**Tc**, $\mathbf{\Delta}_p$, $\mathbf{D}_n$) would correspond to Shannon $H = -\Sigma_{j=1}^{p} f_j \log(f_j)$. However, for the sake of simplicity, the Shannon index will not be considered further in this paper.

Explaining the Occurrences by Environment

Analysis of the statistical triplet based on **Tc** is of no interest by itself except that it links total inertia to a diversity index through the choice of species weights. Various meaningful analyses may nevertheless be conducted in the general framework of Principal Component Analysis on Instrumental Variables (PCAIV; Rao 1964), more widely known as the Redundancy Analysis (RDA; Wollenberg 1977), by analyzing **Tc** with respect to external variables that provide biological, geographical, or environmental information for each occurrence.

Let us consider a table **X**, with n rows (occurrences) and m columns (external variables). Each row of **X** represents a vector of environmental data facing a single individual in the taxonomic table. Multiple linear regressions of each column in **Tc** (response variables) on all columns of **X** (explanatory variables) result in a table $\mathbf{Tc}_\mathbf{X}$ containing the fitted values and a table $\mathbf{Tc}_{|\mathbf{X}}$ containing the residuals $\mathbf{Tc} - \mathbf{Tc}_\mathbf{X}$. The statistical triplet (**Tc**, $\mathbf{\Delta}_p$, $\mathbf{D}_n$) can thus be broken down into two additive parts providing two new statistical triplets (see Appendix B): ($\mathbf{Tc}_\mathbf{X}$, $\mathbf{\Delta}_p$, $\mathbf{D}_n$), which concerns the part of **Tc** explained by the environment and whose eigenanalysis corresponds to PCAIV, and ($\mathbf{Tc}_{|\mathbf{X}}$, $\mathbf{\Delta}_p$, $\mathbf{D}_n$), which concerns the part independent from **X** and whose eigenanalysis corresponds to Orthogonal PCAIV (Sabatier 1984, Sabatier et al. 1989), also known as partial RDA (Davies and Tso 1982). This decomposition divides the total inertia of table **Tc** into a part explained and a part unexplained by the variables contained in table **X**:

$$\text{Iner}(\mathbf{Tc}, \Delta_p, \mathbf{D}_n) = \text{Iner}(\mathbf{Tc}_\mathbf{X}, \Delta_p, \mathbf{D}_n) + \text{Iner}(\mathbf{Tc}_{|\mathbf{X}}, \Delta_p, \mathbf{D}_n).$$

Partitioning the Occurrences: Measurements of α and β Diversity

Let us consider for instance a qualitative variable that partitions the occurrences into m classes. Such a variable may be truly environmental (e.g., soil classes) or more instrumental (e.g., sampling units or relevés). Table **X**, with n rows (occurrences) and m columns (classes of the qualitative variable), contains dummy variables, such that

$$\mathbf{X}_{(n\times m)} = [x_{ik}] = \begin{cases} 1 & \text{if the } i\text{th occurrence belongs} \\ & \quad \text{to the } k\text{th class} \\ 0 & \text{otherwise.} \end{cases}$$

As **X** contains dummy variables, $\mathbf{Tc}_\mathbf{X}$ contains the average species profiles for the classes (i.e., $f_{j/k} - f_j$, where $f_{j/k}$ is the relative frequency of the jth species in the kth class). Consequently, the rows in this table are identical for all occurrences that belong to the same class.

The statistical triplets ($\mathbf{Tc}_\mathbf{X}$, $\mathbf{\Delta}_p$, $\mathbf{D}_n$) and ($\mathbf{Tc}_{|\mathbf{X}}$, $\mathbf{\Delta}_p$, $\mathbf{D}_n$) can now be analyzed using alternatively one of the two matrices of column weights given by $\mathbf{\Delta}_p = \mathbf{D}_p^{-1}$ or $\mathbf{\Delta}_p = \mathbf{I}_p$. These analyses, which are identical to PCAIV and Orthogonal PCAIV with qualitative instrumental variables, are called between- and within-class analyses in the context of partitioned occurrence data (Dolédec and Chessel 1989). The between-class inertia (I_B) is the part of the total inertia of table **Tc** explained by **X**, while the within-class inertia (I_W) is the part of the total inertia of table **Tc** unexplained by **X**. Hence, partitioning the occurrences according to a qualitative ex-

TABLE 1. Key for the soil classes.

Code	Description	fr†
DVD	deep vertical drainage	0.312
Alt	weathered material at <1.2 m depth	0.194
SLD1	superficial lateral drainage with ''dry to the touch'' character (DC) between 1.0 and 1.2 m depth	0.0604
SLD2	superficial lateral drainage with DC at <1.0 m depth	0.0814
DhS	downhill transformed hydromorphic system	0.139
DhS + DC	DhS with DC at <1.2 m depth	0.0446
UhS	uphill transformed hydromorphic system	0.0656
UhS + DC	UhS with DC at <1.2 m depth	0.0472
SH‡	prolonged surface water saturation	0.0551

Note: The soil sequence from the initial ferralitic cover (DVD) up to the transformed hydromorphic systems (DhS and UhS) characterized a weathering transformation process under mechanical erosion.

† Relative frequency of trees ≥50 cm dbh per soil type in a 10-ha rainforest plot in French Guiana.

‡ SH corresponds to the periodically flooded bottomlands and is relatively independent of the weathering process (see Sabatier et al. [1997] for more details).

ternal variable gives the following decomposition: $I_T = I_B + I_W$, where I_T measures the total species diversity of the community (i.e., the γ diversity of Whittaker 1972), broken down into between (β diversity) and within (α diversity) additive components (Lande 1996).

It can be demonstrated (see Appendix C) that the eigenanalysis of ($\mathbf{Tc}_\mathbf{X}$, $\mathbf{\Delta}_p$, $\mathbf{D}_n$) is a Correspondence Analysis (CA) when $\mathbf{\Delta}_p = \mathbf{D}_p^{-1}$, and a Non-Symmetric Correspondence Analysis (NSCA) when $\mathbf{\Delta}_p = \mathbf{I}_p$. It follows that the eigenanalyses of ($\mathbf{Tc}$, $\mathbf{\Delta}_p$, $\mathbf{D}_n$), ($\mathbf{Tc}_\mathbf{X}$, $\mathbf{\Delta}_p$, $\mathbf{D}_n$), and ($\mathbf{Tc}_{|\mathbf{X}}$, $\mathbf{\Delta}_p$, $\mathbf{D}_n$) compute I_T, I_B, and I_W, respectively, either in the χ^2 ($\mathbf{\Delta}_p = \mathbf{D}_p^{-1}$) or the Euclidian metric ($\mathbf{\Delta}_p = \mathbf{I}_p$) (i.e., measuring species diversity by species richness or Simpson index). For the sake of simplicity, the text below will refer to a CA-richness strategy as an analysis that uses $\mathbf{D}_p^{-1}$ as species weights, and a NSCA-Simpson strategy as an analysis that uses $\mathbf{I}_p$ as species weights.

The weighted average α diversity within the classes of $\mathbf{X}$, is $I_W = \Sigma_{k=1}^m f_k I_k$, where I_k is the within-class inertia of k (Lande 1996). Here it is noteworthy that the NSCA-Simpson strategy gives $I_k = \Sigma_{j=1}^p f_{j/k}(1 - f_{j/k}) = 1 - \Sigma_{j=1}^p f_{j/k}^2$, which is exactly Simpson diversity within the kth class of $\mathbf{X}$. Conversely, the CA-richness strategy gives $I_k = \Sigma_{j=1}^p 1/f_j \times f_{j/k}(1 - f_{j/k}) = \Sigma_{j=1}^p (f_{j/k} - f_{j/k}^2)/f_j$, and therefore more weight to the species that are rare in the overall community but abundant in class k, and less weight to the species that are common in the overall community but rare in class k.

Application to Forest Ecological Data

The data were obtained in a 10-ha forest plot at the Piste de St-Elie station in the lowland rainforest of French Guiana (transect B in Sabatier et al. 1997). In this paper we have considered all the trees (381 individuals) with a diameter at breast height (dbh) ≥50 cm. These corresponded to 113 species (species nomenclature follows Boggan et al. 1997) of which 2 comprised >20 individuals and 97 <5 individuals. Taxonomic data were arranged as a complete binary species occurrence table, with 381 tree occurrences as rows and 113 species as columns, secondarily centered by columns (species) in table **Tc**.

Break down of diversity with respect to soil classes

A soil map of the plot (Guillaume 1992) was used to allocate each tree to one of nine soil classes determined in relation to both a weathering transformation sequence of the initial ferralitic cover and a gradient of increasing hydromorphy (Table 1). Soil data were arranged in a complete binary table **S**, with 381 tree occurrences as rows and nine soil classes as columns.

We compared both the CA-richness and the NSCA-Simpson strategies applied to the centered-by-species occurrence table **Tc**, and to the approximated and residual tables, $\mathbf{Tc}_\mathbf{S}$ and $\mathbf{Tc}_{|\mathbf{S}}$, obtained from multiple linear regressions of the columns in **Tc** on all columns of table **S** containing the soil variables. The results are given in Table 2. The inertia of **Tc** measures the total species diversity of the community (i.e., species richness −1 = 112 from the CA-richness strategy and Simpson $D = 0.9331$ from the NSCA-Simpson strategy). The inertias of $\mathbf{Tc}_\mathbf{S}$ and $\mathbf{Tc}_{|\mathbf{S}}$ measure α and β diversity defined by the soil classes.

TABLE 2. Inertia decomposition of a centered-by-species occurrence table (**Tc**, with 381 occurrences of trees ≥50 cm dbh as rows and 113 species as columns) analyzed with respect to nine soil types (table **S** of qualitative explanatory variables) in a 10-ha rainforest plot in French Guiana, using the CA-richness ($\mathbf{D}_p^{-1}$ as species weights) and the NSCA-Simpson strategy ($\mathbf{I}_p$ as species weights).

Ordination strategy	Total species diversity $I(\mathbf{Tc})$	Diversity explained by soil $I(\mathbf{Tc}_\mathbf{S})$	Diversity unexplained by soil $I(\mathbf{Tc}_{\|\mathbf{S}})$	$\frac{I(\mathbf{Tc}_\mathbf{S})}{I(\mathbf{Tc})}$
CA-richness	112	2.33 NS	109.67	2.08%
NSCA-Simpson	0.9331	0.0376 ***	0.896	4.02%

Note: The last column gives the proportion of the total species diversity explained by the soil variables.

*** $P < 0.001$; NS = not significant. Row permutation tests (Manly 1991).

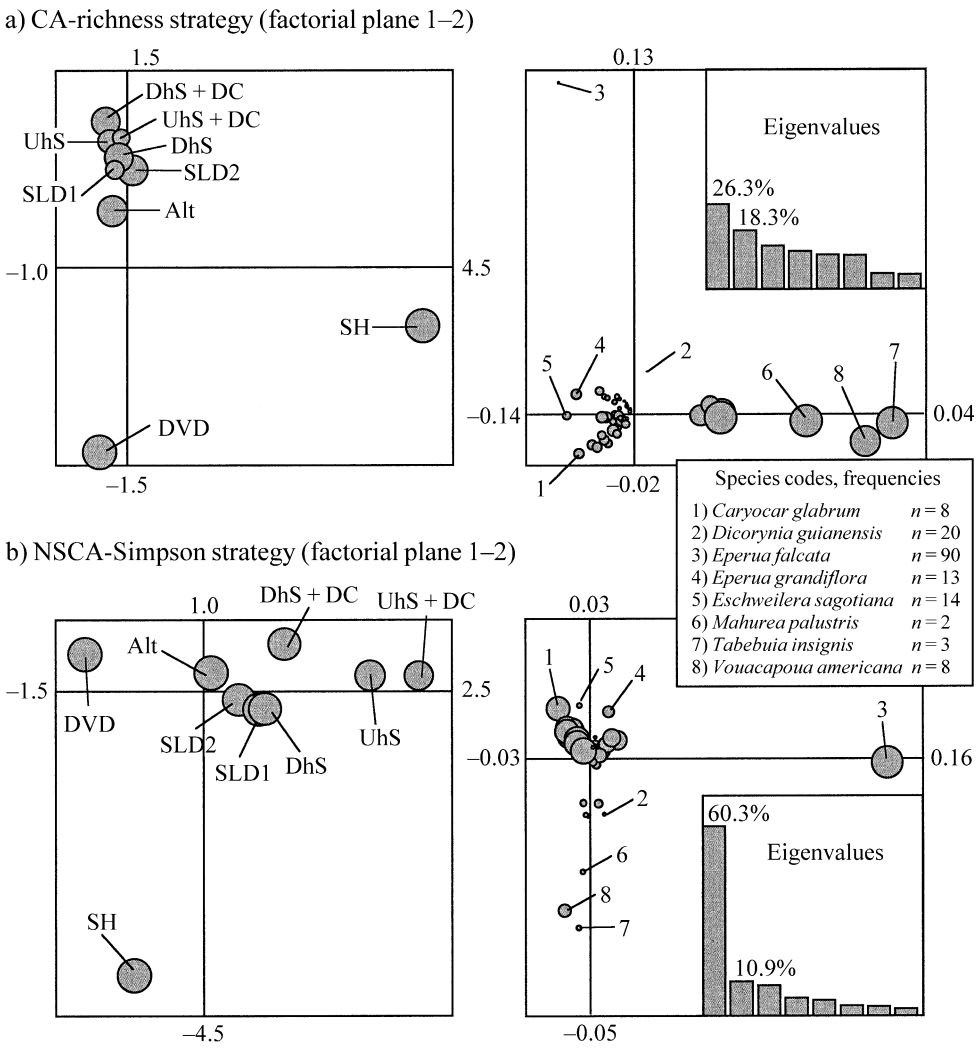

FIG. 2. Ordination of the approximated table $\mathbf{Tc}_S$ (with 381 occurrences of trees ≥50 cm dbh as rows and 113 species as columns) derived from the analysis of a centered-by-columns occurrence table (**Tc**) with respect to nine qualitative explanatory soil variables (table **S**) in a 10-ha rainforest plot in French Guiana. (a) CA-richness strategy ($\mathbf{D}_p^{-1}$ as species weights); (b) NSCA-Simpson strategy ($\mathbf{I}_p$ as species weights). The soil classes and species are positioned by averaging at the weighted mean of their occurrences. Gray circles denote within-class α diversity of the soil classes (left figures) or species relative contributions to axis 1 (right figures). The key for the soil classes is given in Table 1.

In both analyses, the ratio $I(\mathbf{Tc}_S)/I(\mathbf{Tc})$ (i.e., the proportion of the total species diversity of the community explained by the soil classes) was low (<5%) but highly statistically significant using the NSCA-Simpson strategy (row permutation test: $P < 0.001$). In this very diverse forest characterized by an important cortège of scarce species, the NSCA-Simpson strategy places emphasis on the most abundant species and allowed the soil classes to explain a portion of the total species diversity that was about twofold that given by the CA-richness strategy (4.02% vs. 2.08%). This led to a more accurate characterization of the floristic structure on the two first factorial axes in the former case than in the later (axes 1 and 2 correspond to 71.25% vs. 44.62% of the inertia of table $\mathbf{Tc}_S$).

Changing species weights also reversed the hierarchy between the main axes obtained through the analysis of $\mathbf{Tc}_S$ (Fig. 2). Indeed, the first axis resulting from the CA-richness strategy (26.3% of $\mathbf{Tc}_S$ inertia) underlined the originality of the periodically flooded bottomlands (SH) while the second axis (18.3% of $\mathbf{Tc}_S$ inertia) contrasted the ferralitic soils (DVD) with the weathered soil classes (Fig. 2a). On the other hand, the NSCA-Simpson strategy (Fig. 2b) yielded a very prominent axis 1 (60.3% of $\mathbf{Tc}_S$ inertia) expressing more legibly the sequence of soil weathering that led from untransformed DVD to highly weathered soil classes (UhS, UhS + DC). It is noteworthy that the α diversity of the soil classes (represented by gray circles in Fig. 2a and 2b, left side) displayed more variations between

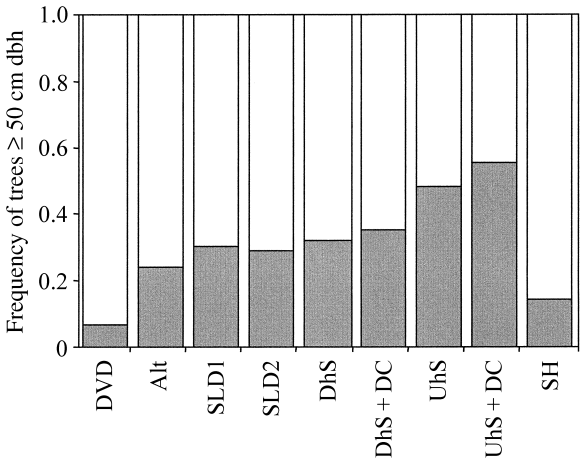

FIG. 3. Relative frequency of trees ≥50 cm dbh belonging to *Eperua falcata* (gray area) and to other species (white area) within nine soil classes in a 10-ha rainforest plot in French Guiana. The key for the soil classes is given in Table 1.

soil classes with the CA-richness than with the NSCA-Simpson strategy. In particular, SLD1 and UhS + DC exhibited very low α diversity in the CA-richness strategy, indicating that they harbored fewer rare species than the other soil classes. The relative species loadings on axis 1 (represented by gray circles in Fig. 2a and 2b, right side) highlighted the fact that influential species in the CA-richness strategy were mainly rare species confined to SH. By contrast, more common species played a prominent role in the NSCA-Simpson strategy. In particular, the distribution of *Eperua falcata*, an abundant and ubiquitous species in this region of French Guiana, corresponded in the study plot to the weathering soil sequence. This points towards a floristic pattern of a broader significance than the sole specificity of the flooded locations (Fig. 3).

Partitioning the occurrences into quadrats

In order to illustrate the possible extension of the method to classical ecological relevés, we partitioned the plot into contiguous quadrats. These quadrats may be considered to be homologous to sampling units of a limited area that are often used to collect individual trees in forest inventories. Three partitions with quadrat sizes of 20 m × 20 m, 25 m × 25 m, and 50 m × 25 m were used. Each gave a complete binary table **Q** of explanatory variables, with 381 tree occurrences as rows and as many columns as quadrats (i.e., 250, 160, and 80 columns for the three partitions). The results of the analysis of **Tc**, $\mathbf{Tc}_{\mathbf{Q}}$, and $\mathbf{Tc}_{|\mathbf{Q}}$ using the CA-richness and the NSCA-Simpson strategies are given in Table 3.

The proportion of the total species diversity of the community explained by the partition into quadrats was high and very similar with both strategies: $I(\mathbf{Tc}_{\mathbf{Q}})/I(\mathbf{Tc})$ ranged from 21.10% to 54.13% using the CA-richness strategy and from 22.90% to 54.11% using the NSCA-Simpson strategy. Logically, this proportion always decreased with quadrat area since α diversity increased with quadrat size. However, $I(\mathbf{Tc}_{\mathbf{Q}})$ was only statistically significant for quadrats of 50 m × 25 m and 25 m × 25 m using the NSCA-Simpson strategy (row permutation tests: $P < 0.01$). Hence, the only significant feature of β diversity among quadrats consisted of variations in the abundance of the most common species, and required a quadrat area in excess of 25 m × 25 m.

Factorial axes were computed from the $\mathbf{Tc}_{\mathbf{Q}}$ table approximated by the partition into 80 quadrats measuring 50 m × 25 m (Fig. 4). Although ordinations were not constrained by the soil variables, the factorial planes, when featuring projections of soil classes, were remarkably similar to those obtained from the analyses of $\mathbf{Tc}_{\mathbf{S}}$ (see Fig. 2 and Fig. 4). This was also the case when using a partition into smaller quadrats (not

TABLE 3. Inertia decomposition of a centered-by-species occurrence table (**Tc**, with 381 occurrences of trees ≥50 cm dbh as rows and 113 species as columns) analyzed with respect to explanatory spatial variables (table **Q**) in a 10-ha rainforest plot in French Guiana, using the CA-richness ($\mathbf{D}_p^{-1}$ as species weights) and the NSCA-Simpson strategy ($\mathbf{I}_p$ as species weights).

Ordination strategy	Total species diversity $I(\mathbf{Tc})$	Diversity explained by quadrats $I(\mathbf{Tc}_{\mathbf{Q}})$	Diversity unexplained by quadrats $I(\mathbf{Tc}_{\|\mathbf{Q}})$	$I(\mathbf{Tc}_{\mathbf{Q}})/I(\mathbf{Tc})$†
80 quadrats of 50 m × 25 m				
CA-richness	112	23.63 NS	88.37	21.10%
NSCA-Simpson	0.9331	0.2137 **	0.7194	22.90%
160 quadrats of 25 m × 25 m				
CA-richness	112	44.90 NS	67.10	40.09%
NSCA-Simpson	0.9331	0.3847 **	0.5483	41.23%
250 quadrats of 20 m × 20 m				
CA-richness	112	60.63 NS	51.37	54.13%
NSCA-Simpson	0.9331	0.5049 NS	0.4282	54.11%

** $P < 0.01$; NS = not significant. Row permutation tests (Manly 1991).
† The proportion of the total species diversity explained by quadrats.

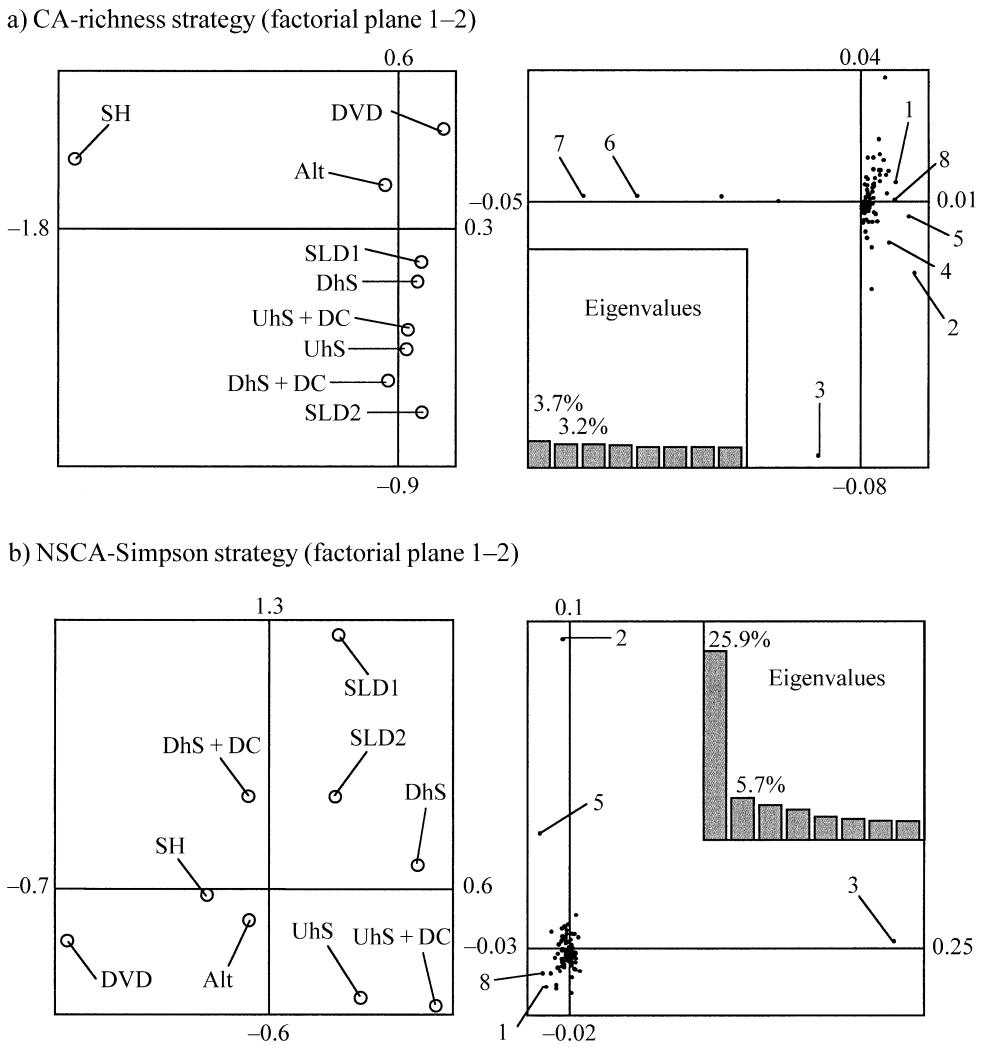

FIG. 4. Ordination of the approximated table $\mathbf{Tc_Q}$ (with 381 occurrences of trees ≥50 cm dbh as rows and 113 species as columns) derived from the analysis of a centered-by-columns occurrence table (**Tc**) with respect to 80 qualitative explanatory spatial variables (table **Q** of quadrats of 50 m × 25 m) in a 10-ha rainforest plot in French Guiana. (a) CA-richness strategy ($\mathbf{D}_p^{-1}$ as species weights); (b) NSCA-Simpson strategy ($\mathbf{I}_p$ as species weights). The soil classes and species are positioned by averaging at the weighted mean of their occurrences. The key for the soil classes is given in Table 1. The key for species is the same as in Fig. 2.

shown). Hence, despite $I(\mathbf{Tc_Q}) > I(\mathbf{Tc_S})$, the analysis of $\mathbf{Tc_Q}$ was not more informative in terms of floristic structures. The distribution of eigenvalues demonstrated, moreover, that the CA-richness approach was far less efficient at extracting meaningful patterns from $\mathbf{Tc_Q}$ than from $\mathbf{Tc_S}$, though the NSCA-Simpson strategy performed fairly well, even on $\mathbf{Tc_Q}$. This latter analysis yielded the most notable departure from Fig. 2 through its second axis (Fig. 4b) which contrasted upland quadrats (negative part) from slopes and lowlands quadrats (positive part) on the basis of the abundance of the second most common species, *Dicorynia guianensis*, whose aggregated spatial distribution (Collinet 1997) was more largely explained by the quadrat partition (20.43%) than the soil classes (5.15%).

Comparison with classical ordinations based on quadrats

We demonstrated (see Appendix C) that analyzing $\mathbf{Tc_Q}$ through either CA or NSCA is strictly equivalent to performing these analyses on a more classical quadrats-by-species contingency table. Furthermore, such a contingency table can be coupled with an ancillary table containing soil information, via Canonical Correspondence Analysis (CCA; Ter Braak 1986) or via a form of Redundancy Analysis (RDA) consistent with NSCA. A table conveying soil information can be constructed from our data by assigning to each quadrat its modal soil type (binary table) or the relative importance of soil types within it (quantitative table based on occurrences). By accepting such a loss of information we

TABLE 4. Inertia decomposition of a quadrats-by-species contingency table (with 80 quadrats of 50 m × 25 m as rows and 113 species as columns) analyzed with respect to an explanatory soil table (with 80 quadrats as rows and nine soil classes as columns) in a 10-ha rainforest plot in French Guiana, using the CA-richness and NSCA-Simpson strategies.

CA-richness strategy			NSCA-Simpson strategy		
	CCA			RDA	
CA	Bin	Quant	NSCA	Bin	Quant
23.63	2.47 (10.43%)	2.57 (10.89%)	0.2137	0.0300 (14.05%)	0.0365 (17.08%)

Notes: The soil table is either a binary table assigning to each quadrat its modal soil type (Bin) or a quantitative table assigning to each quadrat the occurrence relative frequency per soil type (Quant). The proportion of explained inertia is given in parentheses. Abbreviations are defined as follows: CA = Correspondence Analysis; NSCA = Non-Symmetric Correspondence Analysis; CCA = Canonical Correspondence Analysis; RDA = Redundancy Analysis.

are now back in a framework that will appear more usual to most vegetation scientists, and this yields the inertia decomposition presented in Table 4.

The proportion of the inertia explained by the soil variables was 10.43% (binary table) and 10.89% (quantitative table) from CCA, and 14.05% (binary table) and 17.08% (quantitative table) from RDA. The explanatory power of the soil factor may appear more convincing from these values than from the results given in Table 2. However, the proportion of the inertia explained by the soil table in CCA and RDA conveys the same intuitive meaning as the proportion of the inertia of the approximation by **S** of the approximation of **Tc** by **Q** (i.e., of $\mathbf{Tc}_{\mathbf{QonS}}$ obtained from the multiple linear regressions of the columns in $\mathbf{Tc}_{\mathbf{Q}}$ on all columns of **S**). (There is no strict mathematical equivalence here since there are at least two ways to define $\mathbf{Tc}_{\mathbf{QonS}}$ from CCA; Méot et al. 1998). Anyhow, it is clear that CCA and RDA, which start from the quadrats-by-species table, measure β diversity between quadrats (i.e., $I(\mathbf{Tc}_{\mathbf{Q}})$, instead of the total species diversity of the community), and thus ignore the α diversity within quadrats. One advantage of the occurrence-based approach, even for species-by-relevés data, is therefore to provide a fully explicit breakdown of the total floristic diversity of a community, $I(\mathbf{Tc})$, as quantified by usual indices.

DISCUSSION

We have presented two main strategies for the analysis of species occurrence data with respect to environmental or instrumental information. Each strategy encompasses an ordination technique (Correspondence Analysis or CA vs. Non-Symmetric Correspondence Analysis or NSCA) and a consistent measurement of species diversity (species richness vs. Simpson diversity). The CA-richness strategy places emphasis on scarce species while the NSCA-Simpson strategy primarily relies on abundant species. Both strategies enable successive decompositions of the total inertia of a species occurrence table, and thus of the total diversity of the community, between fractions that are explained by certain environmental variables and fractions that are not. A similar scheme of successive decompositions based on CA has already been proposed by Ter Braak (1986, 1987, 1988) for taxonomic data originating from field plots or relevés. This is usually referred to as the Canonical Correspondence Analysis (CCA), though Correspondence Analysis on Instrumental Variables (CAIV) may be preferable (Sabatier et al. 1989; Dray 2001 [available online, see footnote 5]). However, this approach does not rely directly on the floristic diversity of a community as quantified by usual indices.

The paper presented here provides a wider scope based on the notion of species occurrences. Indeed, through the choice of species weighting, our approach allowed us to equate the taxonomic table inertia with common diversity indices. It also permits the use of a wide range of sampling schemes as individuals enumerated in independent field plots (e.g., quadrats, relevés) can be considered as particular types of occurrence data grouped according to an external qualitative spatial variable that codes the plots. This formulation, though unusual, is strictly equivalent to the well-known species-by-relevés table usually considered through CA or CCA when additional environmental variables are available. Subsequent analyses with respect to environmental variables can then be carried out on the fraction of the total inertia explained by the between-relevés analysis (diversity within relevés is ignored as in CCA). This is fully equivalent to directly starting the process from the species-by-relevés table, as via CCA. Even at this stage, the choice between the CA-richness and the NSCA-Simpson strategy remains since CCA can be paralleled by a Redundancy Analysis (RDA) using uniform species weights. Each strategy has strengths and weaknesses. Hence, the choice between the two depends on community characteristics and on the nature of the information to be analyzed.

In fact, a species occurrence table is a unifying mathematical formulation that can be constructed from very diverse field data. Occurrences may correspond to species observations made by successive generations of field investigators without any consistent sampling scheme (e.g., data from museum or herbarium collections; for an example see Gimaret-Carpentier et al. 1998*a*). Exhaustive sampling of large field plots, such as the 10-ha plot presented in this paper, also produces a table for which an occurrence corresponds to an individual tree. The last kind of occurrence data derives from the widely used sampling through plots or relevés within which individuals-by-species are enumerated.

The relevance of the CA-richness strategy vs. NSCA-Simpson strategy greatly depends on the aspect of the plant (or animal) community that one may want to emphasize and also on the broad characteristics of the

community under study. For instance, in very rich vegetation types such as the tropical rainforest, an estimation of species richness can neither be precise nor unbiased on a local scale: all scarce species that can grow in a given context would not be present in a sampling plot of limited area (Condit et al. 1996, Gimaret-Carpentier et al. 1998*b*). Consequently, the relative abundance of scarce species and the total inertia of the CA-richness strategy are ill defined. Hence, the NSCA-Simpson strategy is likely to provide more robust species–environment relationships than the CA-richness approach. In other words, scarce species are less reliable than abundant species when partitioning local diversity into α and β components. Analyzing occurrences of scarce species is nevertheless useful when assessing and comparing γ diversity on a regional scale. This can be done by compiling existing information from local ecological studies as well as large-scale botanical surveys. At this level, the heterogeneity of the data casts doubts on the relevance of the NSCA-Simpson strategy since in the absence of a consistent sampling design there is no guarantee that the most frequent species in the data set are also the most frequent in the region under consideration.

Indeed, an important criterion when choosing species weightings is the reliability of the information conveyed by the absence of a given species in a certain environmental situation (e.g., soil class or geographical quadrat). In the absence of a thorough sampling scheme, such information is dubious since some species could have been missed. Furthermore, the relative abundance of species in the data set is likely to be biased with respect to their relative abundance in the field. The use of the χ^2 metric, through species weights given by $\mathbf{D}_p^{-1}$, is hence highly reasonable and the CA-richness strategy, which guarantees that the species absent from a given environmental modality do not contribute to its ordination score (Legendre and Legendre 1998), is obviously to be preferred. On the other hand, several sampling schemes (either random or systematic; Cochran 1977) may warrant a fair approximation of the species relative abundance in the community under study. This is also true with exhaustive sampling within a large plot, at least for the plot itself. In such cases, the absence of a species in a given environmental context is truly informative and the closer the relative frequency of a species to 0.5 the more information it conveys. This is fully consistent with a measurement of diversity through the Simpson index, as well as with the use of the Euclidian metric (i.e., of uniform species weights given by $\mathbf{I}_p$ instead of $\mathbf{D}_p^{-1}$). Hence, the NSCA-Simpson strategy appears to be more relevant than the CA-richness strategy, and is likely to provide greater insight into the data set.

The link between the sampling scheme and the choice of an ordination technique has already been emphasized with respect to the weighting of sampling units and to its incidence on species niche measurement by Dolédec et al. (2000). Thus, it has been recognized that CA (or CCA) can be used to compensate for unequal sampling efforts, while data from equitable sampling efforts can be analyzed by other ordination techniques, among which stands NSCA on species profiles.

Although this paper grew out of efforts to analyze floristic data based on species distributions, it should be emphasized that the overall principles apply to any kind of organism and/or taxonomic level.

Computation

All analyses presented in this paper have been performed using the ADE-4 statistical package (Thioulouse et al. 1997), which is available online.[6] A user's manual to CA-richness and NSCA-Simpson strategies is available (see Appendix D).

Acknowledgments

This paper stems from important insights provided by D. Chessel through informal discussions and also from documentation accompanying ADE-4 software. This work was partly supported by the GIS Silvolab-Guyane through the AME-Project directed by D. Sabatier. The authors are very grateful to all the botanists and soil scientists who participated in the enumeration and mapping of the 10-ha plot at Piste de St-Elie. They also wish to thank their colleagues, and particularly D. W. Roberts, P. Legendre, B. McCune, and two anonymous referees who made very useful comments on the manuscript.

Literature Cited

Begon, M., J. L. Harper, and C. R. Townsend. 1996. Ecology: individuals, populations and communities. Blackwell Science, Oxford, UK.

Boggan, J., V. Funk, C. Kelloff, M. Hoff, G. Cremers, and C. Feuillet. 1997. Checklist of the plants of the Guianas (Guyana, Surinam, French Guiana). National Museum of Natural History, Smithsonian Institution, Washington, D.C., USA.

Cochran, W. G. 1977. Sampling techniques. Wiley, New York, New York, USA.

Collinet, F. 1997. Essai de regroupement des principales espèces structurantes d'une forêt dense humide d'après l'analyse de leur répartition spatiale (Forêt de Paracou, Guyane). Thèse de Doctorat. Université Claude Bernard, Lyon, France.

Condit, R., S. P. Hubell, J. V. LaFrankie, R. Sukumar, N. Manokaran, R. B. Foster, and P. S. Ashton. 1996. Species–area and species–individual relationships for tropical trees: a comparison of three 50-ha plots. Journal of Ecology **84**: 549–562.

Davies, P. T., and M. K.-S. Tso. 1982. Procedures for reduced-rank regression. Applied Statistics **31**:244–255.

Dolédec, S., and D. Chessel. 1989. Rythmes saisonniers et composantes stationnelles en milieu aquatique. II. Prise en compte et élimination d'effets dans un tableau faunistique. Acta Oecologica **10**:207–232.

Dolédec, S., D. Chessel, and C. Gimaret-Carpentier. 2000. Niche separation in community analysis: a new method. Ecology **81**:2914–2927.

Dolédec, S., D. Chessel, C. J. F. ter Braak, and S. Champely. 1996. Matching species traits to environmental variables: a new three-table ordination method. Environmental and Ecological Statistics **3**:143–166.

[6] URL: ⟨http://pbil.univ-lyon1.fr/ADE-4/⟩

Escoufier, Y. 1987. The duality diagram: a means of better practical applications. Pages 139–156 *in* P. Legendre and L. Legendre, editors. Development in numerical ecology. Springer-Verlag, Berlin, Germany.
Gauch, H. G., Jr. 1973. The relationship between sample similarity and ecological distance. Ecology **54**:618–622.
Gauch, H. G., Jr., and R. H. Whittaker. 1972. Coenocline simulation. Ecology **53**:446–451.
Gimaret-Carpentier, C., D. Chessel, and J.-P. Pascal. 1998*a*. Non-symmetric correspondence analysis: an alternative for species occurrences data. Plant Ecology **138**:97–112.
Gimaret-Carpentier, C., R. Pélissier, J.-P. Pascal, and F. Houllier. 1998*b*. Sampling strategies for the assessment of tree species diversity. Journal of Vegetation Science **9**:161–172.
Greenacre, M. J. 1984. Theory and applications of correspondence analysis. Academic Press, London, UK.
Greenberg, J. H. 1956. The measurement of linguistic diversity. Language **32**:109–115.
Guillaume, J. 1992. Cartographie du sol sous forêt naturelle en Guyane française. Influence des caractères pédologiques sur la structure de la forêt: étude préliminaire. Mémoire de DEA, ENSA, Rennes, France.
Hill, M. O. 1973. Diversity and evenness: a unifying notation and its consequences. Ecology **54**:427–432.
Hill, M. O. 1974. Correspondence analysis: a neglected multivariate method. Journal of the Royal Statistical Society **23**:340–354.
Krebs, C. J. 1994. Ecology: the experimental analysis of distribution and abundance. Harper Collins, New York, New York, USA.
Lande, R. 1996. Statistics and partitioning of species diversity, and similarity among multiple community. Oikos **76**: 5–13.
Lauro, N., and L. D'Ambra. 1984. L'analyse non symétrique des correspondances. Pages 433–446 *in* E. Diday, editor. Data analysis and informatics. III. Elsevier, Amsterdam, The Netherlands.
Legendre, P., and L. Legendre. 1998. Numerical ecology. Elsevier, Amsterdam, The Netherlands.
Magurran, A. E. 1988. Ecological diversity and its measurement. Croom Helm, London, UK.
Manly, B. J. F. 1991. Randomization and Monte Carlo methods in biology. Chapman and Hall, London, UK.
Méot, A., P. Legendre, and D. Borcard. 1998. Partialling out the spatial component of ecological variation: questions and propositions in the linear modelling framework. Environmental and Ecological Statistics **5**:1–27.
Patil, G. P., and C. Taillie. 1982. Diversity as a concept and its measurement. Journal of the American Statistical Association **77**:548–561.
Pélissier, R., S. Dray, and D. Sabatier. 2002. Within-plot relationships between tree species occurrences and hydrological soil constraints: an example in French Guiana investigated through canonical correlation analysis. Plant Ecology **162**, *in press*.
Pielou, E. C. 1969. An introduction to mathematical ecology. John Wiley and Sons, New York, New York, USA.
Rao, C. R. 1964. The use and interpretation of principal component analysis in applied research. Sankhya **A26**:329–357.
Ricklefs, E. 1990. Ecology. W. H. Freeman, New York, New York, USA.
Sabatier, D., M. Grimaldi, M.-F. Prévost, J. Guillaume, M. Godron, M. Dosso, and P. Curmi. 1997. The influence of soil cover organization on the floristic and structural heterogeneity of a Guianan rain forest. Plant Ecology **131**:81–108.
Sabatier, R. 1984. Quelques généralisations de l'analyse en composantes principales de variables instrumentales. Statistique et Analyse de Données **9**:75–103.
Sabatier, R., J.-D. Lebreton, and D. Chessel. 1989. Principal component analysis with instrumental variables as a tool for modelling composition data. Pages 341–352 *in* R. Coppi and S. Bolasco, editors. Multiway data analysis. Elsevier Science, Amsterdam, The Netherlands.
Shannon, C. E. 1948. A mathematical theory of communication. Bell System Technical Journal **27**:379–423.
Simpson, E. H. 1949. Measurement of diversity. Nature **163**: 688.
Ter Braak, C. J. F. 1983. Principal component biplots and alpha and beta diversity. Ecology **64**:454–462.
Ter Braak, C. J. F. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology **67**:1167–1179.
Ter Braak, C. J. F. 1987. The analysis of vegetation–environment relationships by canonical correspondence analysis. Vegetatio **69**:69–77.
Ter Braak, C. J. F. 1988. Partial canonical correspondence analysis. Pages 551–558 *in* H. H. Bock, editor. Classification and related methods of data analysis. North Holland Press, Amsterdam, The Netherlands.
Thioulouse, J., D. Chessel, S. Dolédec, and J.-M. Olivier. 1997. ADE-4: a multivariate analysis and graphical display software. Statistics and Computing **7**:75–83.
Whittaker, C. B. 1972. Evolution and measurement of species diversity. Taxon **21**:213–251.
Wollenberg, A. L. van den. 1977. Redundancy analysis. An alternative for canonical correlation analysis. Psychometrika **42**:207–219.

APPENDIX A

Triplet notation is available in ESA's Electronic Data Archive: *Ecological Archives* E084-006-A1.

APPENDIX B

Triplet decomposition is available in ESA's Electronic Data Archive: *Ecological Archives* E084-006-A2.

APPENDIX C

Relationships between occurrence and contingency tables are available in ESA's Electronic Data Archive: *Ecological Archives* E084-006-A3.

APPENDIX D

A user's manual to CA-richness and NSCA-Simpson strategies is available in ESA's Electronic Data Archive: *Ecological Archives* E084-006-A4.