Environmental and Ecological Statistics **5,** 1–27 (1998)

Partialling out the spatial component of ecological variation: questions and propositions in the linear modelling framework

ALAIN MÉOT[1,2], PIERRE LEGENDRE[1] and DANIEL BORCARD[3]

[1] *Département de sciences biologiques, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Québec H3C 3J7, Canada*
[2] *UPRESA CNRS 6024, 'Laboratoire de Psychologie Sociale de la Cognition'. Université Clermont-Ferrand II. 34, avenue Carnot. 63037 Clermont-Ferrand, France*

First, we formulate some questions posed by the procedure recently proposed by Borcard *et al.* (1992) and Borcard and Legendre (1994) to partition the ecological variation of a community into different portions related to spatial and environmental descriptors. Working separately on the two steps of this procedure – linear modelling and ordinations on modelled tables – allows us to propose different solutions to these questions. These solutions, which use little-known properties of a linear regression model with two additive factors and no interaction, are also adapted to the case of mixed factors (qualitative and quantitative). These properties are presented in the framework of canonical correlation analysis. In particular, they allow us to propose an alternative to partial regression, which avoids confounding. A detailed illustration is presented.

Keywords: canonical correlation analysis, canonical correspondence analysis, ecological models, geographical polynomials, linear model, oribatid mites, spatial patterns, statistical triplet, variation partitioning

1. Introduction

Borcard *et al.* (1992) and Borcard and Legendre (1994) proposed the use of partial canonical correspondence analysis (CCA: ter Braak, 1986, 1988) to partition the ecological variation of a community into different portions related to spatial and environmental descriptors. In their procedure, the total variation of the species table, which is measured by the inertia of a correspondence analysis, is split up into four different quantities: 'pure spatial', 'pure environmental', 'spatial component of environmental influence' and 'undetermined'. Using a polynomial of the geographic coordinates (latitude and longitude) as spatial descriptors only allows capturing large-scale spatial structures; in that case, the 'pure environmental' component can be associated with the more regional or local variation. Alternatively, more local spatial variation can be captured using an autocorrelation model (Legendre and Borcard, 1994).

We would like to draw attention to two of the reasons that make this approach interesting.

1. Description of the spatio-temporal variability associated with communities is an essential complement to more functional approaches (Prodon and Lebreton, 1994). The Borcard *et al.* procedure is one of the very few propositions to explicitly incorporate spatial structures into ecological modelling. Because ecological data generally show strong spatial structuring, the need for this kind of method is great. Univariate tools have been proposed in geography (Cliff and Ord, 1981; Upton and Fingleton, 1985), econometrics (Griffith, 1988) and geostatistics (Isaaks and Srivastava, 1991; Rossi *et al.*, 1992) to describe spatial structures. The multivariate approach is, however, functional in ecology, but methods available for modelling remain extremely underdeveloped, and they often contain too many non-realistic hypotheses and procedures. For example, global tests of significance involving ecological and spatial distance matrices, such as the Mantel or partial Mantel tests (Mantel, 1967; Hubert *et al.*, 1981; Smouse *et al.*, 1986; Legendre and Troussellier, 1988), are just very rough indicators of the patterns structuring species assemblages in space. They must be completed by detailed descriptive analyses. In addition, different patterns can be responsible for a given observed correlogram, and this is even more true in the multivariate than in the univariate setting (e.g. Sokal and Oden, 1978; Legendre and Fortin, 1989).
2. This approach possesses direct links with classical hypotheses and conceptual models concerning the influences of biotic relationships, environmental effects (e.g. May, 1984; Sokal and Oden, 1978) and other factors, such as historical events or random factors and disturbances (e.g. Denslow 1985; Borcard and Legendre, 1994), on community structures. These effects are usually very difficult to separate. Furthermore, critical causal factors that may have been ignored can produce spurious correlations among the factors included in the analysis. Hence, considering an *a priori* spatial structure as a synthetic model for the processes responsible for the spatial variation, as proposed by Borcard *et al.* (1992), can be of great interest for ecological analysis. Conceptually, this practice is clearly linked with the introduction of latent variables in causal models (e.g. Gifi, 1990, p. 48). Classical ordinations are often dominated by ecological gradients linked to environmental factors. Identifying explanatory variables that are independent of these factors to obtain a 'pure environmental component', as in the procedure of Borcard *et al.*, can be very useful for a detailed understanding of more local ecological structures. In most instances, these structures are numerically of minor magnitude but they remain ecologically meaningful.

The partitioning procedure proposed by Borcard *et al.* presents three important problems, however.

- The first concerns the part of the variability which is interpreted as the 'spatial component of the environmental influence'. The proposed measure of this variability is in fact not a fitted variance component, but the difference between the variances explained by two models which have no particular

relationship to one another. As a result, the measured quantity cannot be defined as a variance explained by a linear model of the species by some particular kind of predictor.

- The second question is less important; it concerns the measures of variability associated with the 'pure spatial' and 'pure environmental' components. These two measures are the variances (inertia) explained by two sets of predictors which generally are not linearly independent of one another. As a result, the use of these two quantities (and models) to judge the effects of the environmental or spatial variables alone leads to some confounding between these effects.
- The third problem concerns a question raised by H.J.B. Birks, in the discussion that follows the paper of Borcard and Legendre (1994), about the possibility of analysing the different fractions, using ordination, in order to explore the effects of the different kinds of predictors on the species assemblage patterns. This question is particularly important for the spatial component of environmental variation. As explained above, this component is not dependent on a model, and thus it cannot readily be decomposed into ordination axes.

Hence, the purpose of this paper is threefold.

- We explicitly present the procedure used by Borcard *et al.* in a framework of regression analysis followed by ordinations of the modelled tables. The nature of the different measures and associated models used in the partitioning procedure is clearer in this way, as well as the questions brought up by this procedure.
- Using this framework, we propose complementary solutions to partition the species variation. These solutions consist in the construction and analysis of successive models of the species table using covariables characterized by properties of greater or lesser statistical independence between the spatial and environmental dimensions. They allow us to map the multivariate structures associated with the different factors under consideration. They are particularly efficient to partial out the spatial component of ecological variation.
- We illustrate the new procedures using the Borcard *et al.* oribatid mite data set (Acari, Oribatei).

2. Theory

2.1 *Notations and general considerations*

Let $\mathbf{P}$ be a data table containing the frequencies of t species (in columns) observed over n objects (in rows). These frequencies are written $p_{ij} (i = 1, \ldots, n; j = 1, \ldots, t)$. The row and column margins of $\mathbf{P}$ define two weightings, denoted by $\{p_{i.}\}$ and $\{p_{.j}\}$, for the rows and columns of $\mathbf{P}$. The two diagonal matrices, $\mathbf{D}_i$ and $\mathbf{D}_j$, including on their diagonals the $p_{i.}$ and $p_{.j}$ values, also define two scalar products in the column and row space representations. Finally, $\mathbf{P}^\circ$ is the table made of the elements of the form:

$$(p_{ij} - p_{i.}p_{.j})/p_{i.}p_{.j}.$$

A little-known way to present the correspondence analysis (CA) of **P** is to consider it as the analysis of the statistical triplet $(\mathbf{P}^{\circ}, \mathbf{D}_i, \mathbf{D}_j)$ (see Escoufier 1982, 1987, and the appendix of Chévenet *et al.*, 1994, for a brief outline of the analysis of a statistical triplet; or Dolédec *et al.*, 1996, for a translation in the framework of general singular value decomposition used by Greenacre, 1984). This method is used in the presentation that follows.

All of the approaches used in this paper explore the relationships between a set of t dependent variables, denoted as the columns of a table **Y**, which is derived from table **P**, and a set of k explanatory variables, denoted as the columns of a table **X**. The explanatory variables are defined on the same objects (rows) as **P**. These approaches comprise two steps:

- A multivariate multiple regression of **Y** on **X** is done. This regression uses a weighted least squares criterion involving the weighting $\{p_{i.}\}$. Results of this regression are written out in the table of fitted values $\mathbf{Y}_{\mathbf{X}}$ and in the table of residuals $\mathbf{Y}_{/\mathbf{X}}=(\mathbf{Y}-\mathbf{Y}_{\mathbf{X}})$.
- The triplet $(\mathbf{Y}_{\mathbf{X}}, \mathbf{D}_i, \mathbf{D}_j)$ is then analysed using the particular relationships created in the first step (see Appendix 1 for a brief mathematical outline). The triplet $(\mathbf{Y}_{/\mathbf{X}}, \mathbf{D}_i, \mathbf{D}_j)$ can also be analysed, which makes it possible to explore the remaining structures of **Y**, after the effects of **X** have been eliminated.

When **Y** is equal to $\mathbf{P}^{\circ}$ and **X** to a table of k environmental variables, the analysis of the triplet $(\mathbf{Y}_{\mathbf{X}}, \mathbf{D}_i, \mathbf{D}_j)$ is equivalent to the CCA of **P** constrained by **X** (e.g. Chessel *et al.*, 1987, Lebreton *et al.*, 1991, or ter Braak and Verdonschot, 1995, Box 3, for a brief outline). In that case, the analysis of the triplet $(\mathbf{Y}_{/\mathbf{X}}, \mathbf{D}_i, \mathbf{D}_j)$ gives the non-canonical axes of CCA as can be found by the CANOCO *program.*

Note that generally the columns of **X** have been centred or standardized using the weights $\{p_{i.}\}$ before the start of the analysis. Note also that **X** can contain variables from which the effect of another set of explanatory variables has been removed. In that case, the analysis is equivalent to a partial CCA.

Most of the analyses used in this paper belong to that case. After controlling for the effect of covariables defined by spatial and/or environmental characteristics of the objects, they follow the normal development of CCA. We use the type-2 scaling of CANOCO (ter Braak, 1990), where the columns (species) are at centroids of the rows (sampling sites). This scaling puts the species at the centroids of the sampling sites in the graph. To simplify, we call environmental site scores the sample scores which are linear combinations of environmental variables. We also consider as environmental variable scores the correlations (using weighting $\{p_{i.}\}$) between the environmental variables and the environmental site scores. This procedure allows us to standardize all our analyses.

For a matrix **Y** different from $\mathbf{P}^{\circ}$, the multivariate multiple regression allows us to decompose it into two independent parts: fitted and residuals. As a result, the inertia associated with the triplet $(\mathbf{Y}, \mathbf{D}_i, \mathbf{D}_j)$ is equal to the sum of the inertias associated with the triplets $(\mathbf{Y}_{\mathbf{X}}, \mathbf{D}_i, \mathbf{D}_j)$ and $(\mathbf{Y}_{/\mathbf{X}}, \mathbf{D}_i, \mathbf{D}_j)$. Successive models can be constructed, analysed and compared using this method. The analyses, which are different from CCA in this case, can be done using the general framework of principal component analyses with respect to instrumental variables (Rao, 1964; Lebreton *et al.*, 1991; Sabatier *et al.*, 1989). An outline of the procedures used in this paper is given in Appendix 1.

In this paper, we focus attention onto the particular case where the species data table is suitable for CA. However, as in Borcard *et al.* (1992), other constrained ordination methods can be used for other types of data tables. For example, if linear relationships between species and environmental traits are expected, redundancy analysis can be obtained using the same procedure as above, by changing the species table and the associated scalar products (weightings) to that used in principal component analysis.

2.2 *Questions raised by the Borcard* et al. *procedure*

Let **E** be the table of p environmental predictors and **S** the table of s spatial predictors. Variables are the columns and sampling sites are the rows of these tables. Columns of **E** and **S** have been standardized beforehand, using the weights $\{p_{i.}\}$.

The first step of the Borcard *et al.* analysis is a decomposition of the species variation in two parts. The first is due to the additive effect of the environmental (**E**) and spatial (**S**) predictors. The second is the residual from this regression procedure.

For the whole table of species data, these models can be noted in matrix form as:

$$\mathbf{P}^{\circ} = \mathbf{P_T} + \mathbf{R}$$

where $\mathbf{P_T}$ is the table containing in columns the model (i.e. the fitted values) and **R** is the table containing in columns the residuals. **R** represents the undetermined fraction of Borcard *et al.* (Fig. 1).

The second step of the analysis consists in decomposing the model contained in $\mathbf{P_T}$ into two independent (orthogonal) parts.

Two decompositions are possible at this stage (e.g. Rao and Yanaï, 1979).

- Model 1

  $$\mathbf{P_T} = \mathbf{P_E} + \mathbf{P_{S/E}} \quad \text{and} \quad \mathbf{P}^{\circ} = \mathbf{P_E} + \mathbf{P_{S/E}} + \mathbf{R}$$

  where $\mathbf{P_E}$ is the table containing in columns the regression models of the columns of $\mathbf{P_T}$ on the **E** variables alone and $\mathbf{P_{S/E}}$ contains in columns the residuals from those regressions. Note that because the **E** and **S**/**E** predictors are nested in the $\mathbf{E} + \mathbf{S}$ matrix, these two parts can be directly obtained by the regressions of $\mathbf{P}^{\circ}$ on the environmental predictors (result: $\mathbf{P_E}$) and on the residuals of the regression of the spatial variables on the environmental predictors. These residuals are written in table **S**/**E**. The results of the regressions is $\mathbf{P_{S/E}}$.

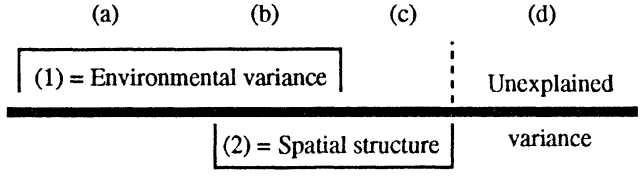

Figure 1. Variation partitioning of the species table, showing that fraction (b) is the intersection of the environmental and spatial components of the species variation. (Modified from Borcard *et al.*, 1992.)

- Model 2:

  $$\mathbf{P_T} = \mathbf{P_S} + \mathbf{P_{E/S}} \quad \text{and} \quad \mathbf{P}^\circ = \mathbf{P_S} + \mathbf{P_{E/S}} + \mathbf{R}$$

  where $\mathbf{P_S}$ is the table containing in columns the regression models of the columns of $\mathbf{P_T}$ on the **S** variables alone and $\mathbf{P_{E/S}}$ contains in columns the residuals from these regressions. Note that in this decomposition, $\mathbf{P_S}$ is also equal to the regressions of $\mathbf{P}^\circ$ on the spatial predictors and $\mathbf{P_{E/S}}$ to the regressions of $\mathbf{P}^\circ$ on the residuals of the regression of the environmental predictors on the spatial variables. These residuals are denoted **E**/**S**.

Models $\mathbf{P_{S/E}}$ and $\mathbf{P_{E/S}}$ are generally not independent of one another because the predictors used to build them are not. As a result, comparisons of the explained to the total variances lead to a confounding of the effects of the **S**/**E** and **E**/**S**-predictors. In classical multivariate analysis of variance, these models are however considered sufficient to analyse the main effect of one factor after the effect of the other has been removed (e.g. Takeuchi *et al.*, 1982, pp. 147–9).

Secondly, using either model 1 or model 2, $\mathbf{P_T}$ is exactly the sum given by two independent models ($\mathbf{P_E} + \mathbf{P_{S/E}}$ for model 1, and $\mathbf{P_S} + \mathbf{P_{E/S}}$ for model 2). This property allows Borcard *et al.* to write in terms of the variabilities taken into account:

variance explained by model 1 = variance explained by model 2

$$<=> (1) + (c) = (2) + (a) \qquad (\mathrm{var}(\mathbf{P_E}) + \mathrm{var}(\mathbf{P_{S/E}}) = \mathrm{var}(\mathbf{P_S}) + \mathrm{var}(\mathbf{P_{E/S}}))$$

$$<=> (1) - (a) = (2) - (c) \qquad (\mathrm{var}(\mathbf{P_E}) - \mathrm{var}(\mathbf{P_{E/S}}) = \mathrm{var}(\mathbf{P_S}) - \mathrm{var}(\mathbf{P_{S/E}}))$$

where (1) and (2) refer to variances identified in Figure 1. Fraction (1) − (a) or (2) − (c) is then considered as the 'spatially structured environmental variation', or part (b) in Fig. 1.

Actually it is not entirely justified to regard this quantity (b) as an amount of variance explained by a linear model of the species by some particular predictors. This is why Borcard *et al.* (1992), Borcard and Legendre (1994) refer to it by the more general term *variation*; they describe fraction (b) as the variation of table **P** that can equally be attributed to variables **E** or **S**. It is computed as the algebraic difference between the variances explained by two different models. Because the **E**/**S** type predictors generally are not strict linear combinations of the environmental variables, the relationships between the two sets of predictors used in these models are complex (but see, for example, Pernin and Pagés, 1988) and can give rise to difficulties of interpretation when comparing the associated explained variances. Note, however, that Whittaker (1984) argues for the use of this kind of index for two reasons: to select relevant explanatory variables in multiple linear regression, and to measure the effective balance of the design, with (b) = 0 for a completely balanced design.

For example, $\mathbf{P}^\circ$ may be better represented on the **E**/**S** type predictors than on the environmental variables. The authors are clearly conscious of this when they say: 'Note that in theory, value (c) [there is an error in the text; they obviously meant fraction (b)] can be negative' (Borcard *et al.*, 1992, p. 1049).

The answer to the question of how to separate the variation which can be attributed to spatially structured environmental factors from that due to non-spatial environmental predictors is thus not completely satisfactory and must be revisited.

2.3 *Other ways of determining spatially structured and non-spatial environmental variation*

Using the previous regression models, it clearly appears that the problem that we have pointed out is linked with the decomposition of the analysis of variance, where the question is how to judge the respective effects of two factors in a simple additive model with no interaction. The problem is, however, different in our case because the factors contain a mixture of qualitative and quantitative variables, and the individual weights are not uniform. Because these two sets of factors are generally not orthogonal (i.e. linearly independent), some confounding can appear when measuring their effects using the sum of squares of the models for the two factors. Hence, it is necessary to use non-standard approaches to carry out the analysis.

Because of the dependency between the two sets of predictors, several solutions can be found to the decomposition. This means that it is necessary to first choose one decomposition of $\mathbf{P_T}$: either $\mathbf{P_E} + \mathbf{P_{S/E}}$, or $\mathbf{P_S} + \mathbf{P_{E/S}}$, and then decompose the first term of one of these models ($\mathbf{P_E}$ or $\mathbf{P_S}$) using the second set of predictors (**S** or **E** types). In this paper we have chosen to focus on the first decomposition, $\mathbf{P_E} + \mathbf{P_{S/E}}$. Because the use of a polynomial of the spatial coordinates is simply a practice allowing one to take the large-scale spatial structure into account, it seems more interesting to decompose the 'environmental effects' first by taking into account the spatial predictors, than the converse. But of course, the procedures proposed here can be used symmetrically for the second decomposition.

Although there are several possible ways to carry out this decomposition, two of them appear to be particularly heuristic.

2.3.1 First method

The simplest way consists of decomposing the $\mathbf{P_E}$ table (whose variance is (a + b) in Fig. 1) as the additive model of the fitted and residual values obtained from the weighted linear regressions (the weights are again the row margins of the **P** table) of the columns of this table on the spatial variables. The models (fitted values), written in the columns of table $\mathbf{P_{EonS}}$, can be seen as the part of the model of $\mathbf{P}^{\circ}$ on **E** that is spatially structured; it plays the role of fraction (b) in Fig. 1. The residuals, denoted by $\mathbf{P_{Eon/S}}$, are the parts of these models which are not spatially structured (somewhat equivalent to fraction (a) of Fig. 1). These two fractions are obviously orthogonal and thus, the sum of the associated variabilities is equal to the variability associated with $\mathbf{P_E}$. Analyses of the triplets $(\mathbf{P_{EonS}}, \mathbf{D}_i, \mathbf{D}_j)$ and $(\mathbf{P_{Eon/S}}, \mathbf{D}_i, \mathbf{D}_j)$ allow an appreciation of the relationships between the species and the environmental factors which may be spatially structured or not. Mapping the row scores leads to an understanding of the large-scale or the more local relationships among sampling sites under these models.

This approach is not completely satisfactory, however, because what is measured in the analysis of $\mathbf{P_E}$ by **S** is the effect of the spatial variables on the model of the species table by the environmental variables. Hence, the fitted values cannot be considered as the direct effects of environmental variables, spatially structured or not. One of the problems is that this procedure will favour the environmental variables that explain the largest part of the variation. Because these variables generally show large-scale patterns corresponding to the gradient, small-scale spatial components may be left to play only a minor role. Hence, variables showing exclusively local or regional variation are

not brought out very efficiently using this procedure. Another problem is that some confounding may appear between the decomposition models of $\mathbf{P_E}$ and the third term given by $\mathbf{P_{S/E}}$. This confounding is possible because the $\mathbf{S/E}$ and $\mathbf{S}$ type predictors are not independent. If the purpose is ordination, it is often not very important and a visual inspection of the factor scores will not allow the detection of this confounding.

2.3.2 Second method

Another interesting but mathematically more complex solution can be used. It rests on a decomposition in different orthogonal parts of the vectorial subspace generated by two different sets of predictors. The mathematical bases of this decomposition have been given by Afriat (1957). For statistical purposes, a clear presentation of this decomposition can be found in Pontier and Pernin (1987) and Pontier *et al.* (1990, Chapter 6, particularly pp. 242–52); it uses some properties of canonical correlation analysis. We present a brief outline of this approach in Appendix 2. Different uses and developments of this decomposition in the case of MANOVA designs with two strict qualitative factors can be found in Pontier and Pernin (1987), Pontier *et al.* (1990), Pernin and Pagés (1988), Yoccoz and Chessel (1988) or Fraile *et al.* (1993).

Canonical correlation analysis searches for linear combinations (canonical variables) of the columns of two tables which bring out the highest correlations between tables, with the constraint that these linear combinations be uncorrelated pairwise (e.g. Gittins, 1985). In the present paper, the two tables subjected to analysis contain the spatial polynomial and the environmental variables. Weighted correlations are used with the row margins of $\mathbf{P}$ as weights. The calculation of the coefficients of the linear combinations (canonical coefficients) issued from the two tables is described in Appendix 2. Some of the canonical variables obtained from this analysis are then used as predictors of matrix $\mathbf{P}$ using CCA. Three different kinds of canonical variables can be recognized (Cailliez and Pagès, 1976, p. 357 *et seq.*; Pontier and Pernin, 1987; Pontier *et al.*, 1990).

- It may happen that two canonical variables, one from each set of predictors, are identical. A pair of canonical variables of this type is hence reduced to a single variable, meaning that there is a perfectly common part to the two tables of predictors. A canonical variable like this would mean that one linear combination of the environmental variables is perfectly spatialized. Canonical variables with this property correspond to eigenvalues equal to unity ($r = \lambda = 1$). They are written in the colums of a table denoted by $\mathbf{E} \cap \mathbf{S}$.
- The second kind of linear combination is commonly encountered in canonical correlation analysis. The two canonical variables of a pair, one in each of the two spaces, differ. Pairs are ordered by successively decreasing correlations; they have eigenvalues between 1 and 0. The canonical variables that are associated with the environmental table display a combination of large-scale and more regional-scale variation, but not solely large-scale or regional-scale. They are written in the columns of a table denoted by $\mathbf{O_S}(\mathbf{E})$.
- The third kind of linear combination is usually of no interest in canonical correlation analysis, because it makes little sense to consider them in that framework. It concerns the remaining pairs of linear combinations, for which correlations are zero; these linear combinations are associated with the null eigenvalues of the

analysis. Because the two matrices to be analysed are not of the same size (Appendix 2), the number of eigenvectors associated with null eigenvalues are not the same on both sides of the analysis. There is also no particular relationship between these eigenvectors for the two matrices. If we only consider the linear combinations (canonical variables) dependent on the environmental table, they are uncorrelated with all the canonical variables associated with positive eigenvalues, and also uncorrelated with all the canonical variables derived from the spatial table. $\mathbf{E} \cap \mathbf{S}^{\perp}$ is the table containing in columns this kind of canonical variable.

Our notation refers to the intersection as the part common to two sets of predictors while $^{\perp}$ expresses independence. Hence, $\mathbf{E} \cap \mathbf{S}$ is the part common to the environmental and spatial spaces ; $\mathbf{E} \cap \mathbf{S}^{\perp}$ is the part of the environmental space which is independent of the spatial space; and $\mathbf{O_S}(\mathbf{E})$ is the part of the environmental space which is neither common nor independent of the spatial space. Since independence refers to orthogonality, $\mathbf{O_S}(\mathbf{E})$ is neither included nor orthogonal to the spatial space; it is thus 'oblique'.

Using these three kinds of canonical variables in turn as predictors for $\mathbf{P}$ allows us to decompose table $\mathbf{P}^{\circ}$ into a sum of five independent models:

$$\begin{aligned}\mathbf{P}^{\circ} &= \mathbf{P_E} + \mathbf{P_{S/E}} + \mathbf{R} \\ &= \mathbf{P_{E\cap S}} + \mathbf{P_{O_S(E)}} + \mathbf{P}^{\perp}_{\mathbf{E}\cap\mathbf{S}} + \mathbf{P_{S/E}} + \mathbf{R}.\end{aligned}$$

A bit more explanation is necessary to really understand the meaning of these predictors and to highlight their differences with the Borcard *et al.* approach described above.

2.3.3 Difference between regression models using the $\mathbf{E}/\mathbf{S}$ *type predictors and canonical variables associated with null eigenvalues*

The aim of regression of the species on the $\mathbf{E}/\mathbf{S}$ type predictors (Borcard *et al.* approach) is to identify the linear combinations of the residuals of the regression models of the environmental variables on the spatial variables that best explain the variance of the species (fraction (a) in Fig. 1). These residuals are what is left of the environmental variables after representing them by the spatial predictors. They are independent of the spatial variables; but generally, they are not linear combinations of the environmental variables.

Using as predictors the third kind of independent variables described above ('the remaining pairs of linear combinations, for which correlations are zero') is equivalent to adding a constraint to the regression on the $\mathbf{E}/\mathbf{S}$ type predictors. These predictors must be independent of the spatial variables, but they also have to represent linear combinations of the environmental variables. The linear combinations found using these variables as predictors are thus also linear combinations of the environmental variables. Because these predictors are nested in the $\mathbf{E}$ type, the explained variability (inertia) is thus naturally comparable with that explained by the environmental predictors; this is not the case of the variability explained by the $\mathbf{E}/\mathbf{S}$ type predictors. These predictors allow one to judge the specific effect of the environmental predictors or, in other words, the variation of table $\mathbf{P}$ which is due to the environmental descriptors that are completely independent of any spatial variables; this is a strong equivalent to fraction (a) of the variation in Borcard *et al.*

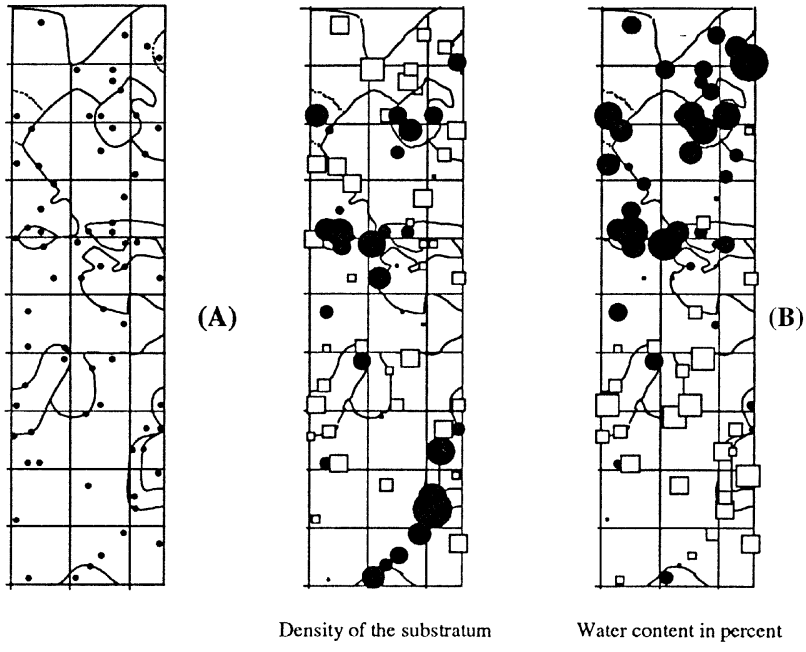

Figure 2. (A) Seventy sampling locations positioned in the Lake Geai plot. Straight lines are 1 m apart. (B) The two quantitative environmental variables, density of the substratum (left) and percentage water content (right), standardized using the row margins of the species table as weights. Squares represent negative values (proportional to size) while circles represent positive values. (C) Distributions of the qualitative environmental variables: substratum, shrub cover and microtopography (from Borcard and Legendre, 1994). These variables (or modalities) will appear in the same order in the analyses.

2.3.4 Using canonical variables associated with unit eigenvalues as predictors

If we used as predictors of **P** the models of the environmental variables obtained by the linear regressions of these variables on the spatial predictors, it is the linear combinations of the spatial variables that best explain the variance of the environmental variables that would be used as the independent variables. The adjusted values of **P** found would not, however, be constrained to be dependent on the environmental variables because the predictors would not be constrained to have that property. Using the canonical variables associated with unit eigenvalues as predictors adds a constraint: the independent variables must be linear combinations of the spatial variables (like in the regression model of **E** on **S**), and they also have to be linear combinations of the environmental variables. The explained variance is thus naturally comparable with that explained by the environmental predictors, which is not the case of the variance explained by the (**E** on **S**) type predictors, or of fraction (b) of Borcard *et al.*

These predictors allow one to judge the common effect of the environmental and spatial factors or, in other words, the part of the variation of the species table which is due to completely spatialized environmental descriptors; this is a strong equivalent to fraction (b). This type of canonical variable is seldom encountered when analysing real data.

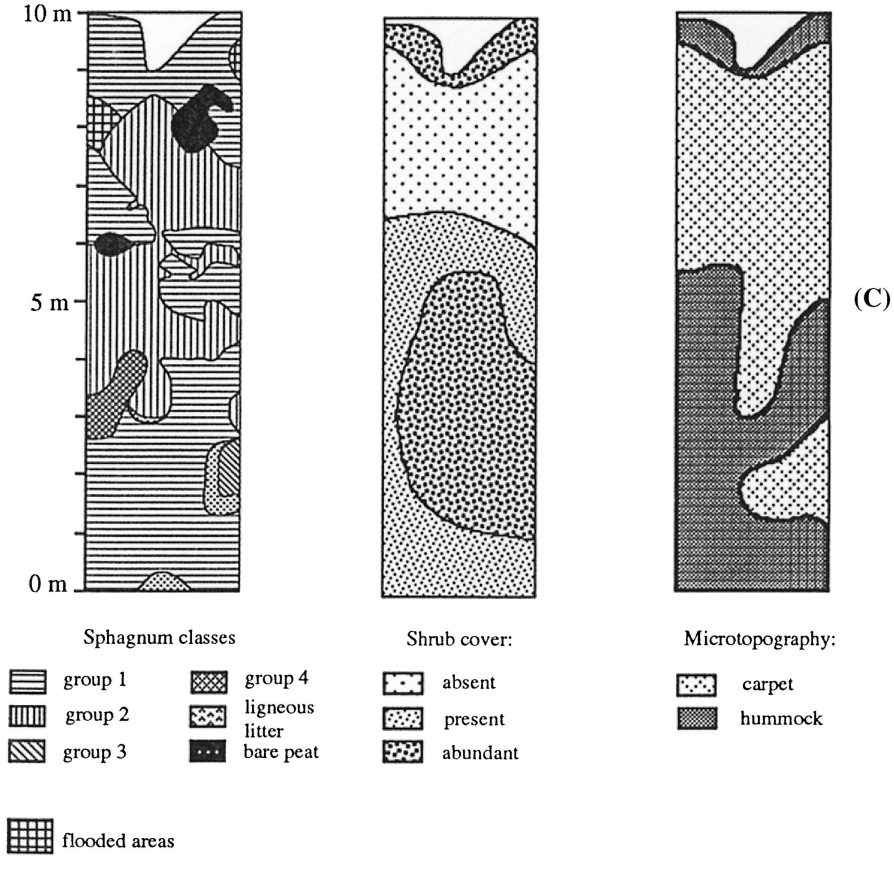

Figure 2. (A) Seventy sampling locations positioned in the Lake Geai plot. Straight lines are 1 m apart. (B) The two quantitative environmental variables, density of the substratum (left) and percentage water content (right), standardized using the row margins of the species table as weights. Squares represent negative values (proportional to size) while circles represent positive values. (C) Distributions of the qualitative environmental variables: substratum, shrub cover and microtopography (from Borcard and Legendre, 1994). These variables (or modalities) will appear in the same order in the analyses.

2.3.5 Using canonical variables associated with eigenvalues between 0 and 1 as predictors

The second kind of predictor is less interesting for our purpose, although commonly found. They represent linear combinations of the environmental variables that are neither completely dependent nor completely independent of the spatial variables. Using them as predictors of the **P** table allows one to judge the effect of the environmental predictors displaying a combination of large-scale and more regional-scale variation. There is no strict equivalent in the Borcard *et al.* method. It is easy to

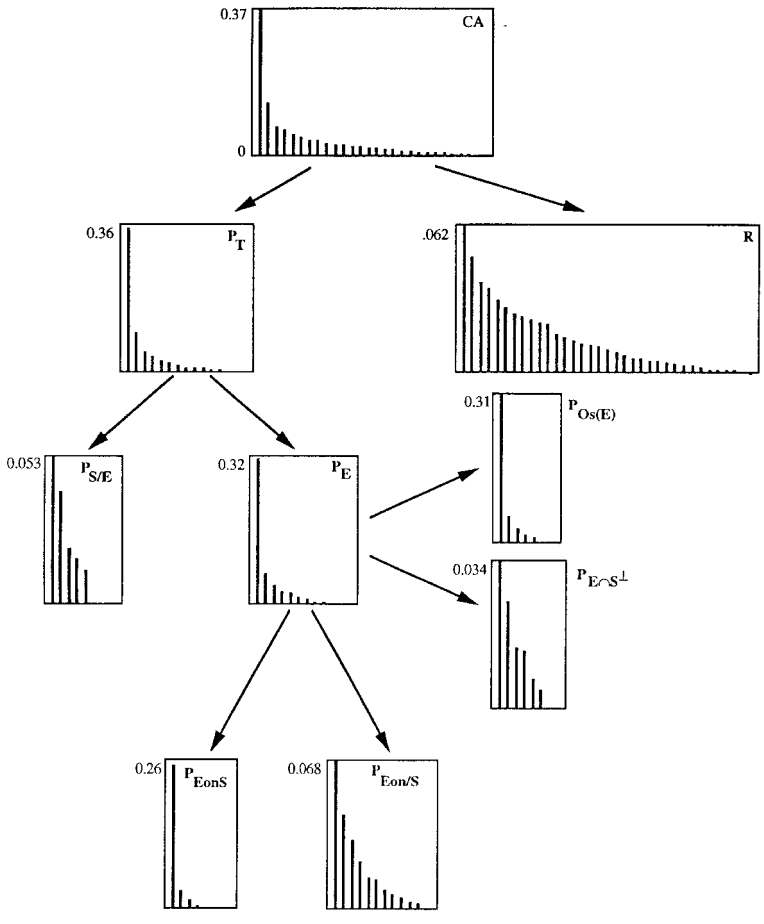

Figure 3. Ranked eigenvalues resulting from the analyses on the different modelled tables. The highest eigenvalue is given at the top left of each plot. The arrows indicate the successive model decompositions in two parts. Models higher up in the hierarchy are obtained by the addition of the models lower down (e.g. $\mathbf{P_E} = \mathbf{P_{EonS}} + \mathbf{P_{Eon/S}} = \mathbf{P_{\bar{E}}^{\perp}} + \mathbf{P_{Os(E)}}$). $\mathbf{P_E}$ has been submitted to the two decompositions (down and left) explained in the text.

understand that they would be the most common type of linear combinations. Some thought should be given to partitioning them between those that are well linked with the spatial variables and those that are not. These developments are, however, beyond the scope of the present paper.

The analyses made in the following sections, using this partitioning method, are simply CCAs of **P** constrained by each kind of predictor. The calculations have been carried out using the ADE software (Thioulouse *et al.*, 1995b) which allows one to calculate the different sets of predictors and to do the CCA. The type-2 scaling of CANOCO was obtained by dividing the species row scores obtained from ADE by the square root of the corresponding eigenvalues.

3. Illustration: the Borcard *et al.* data set on oribatid mites (Acari, Oribatei)

3.1 *The data set*

To illustrate their procedure, Borcard *et al.* have extensively used a particular data set concerning the spatial distribution of oribatid mites (Acari, Oribatei) on the southern shore of a small bog lake, Lake Geai, on the territory of the Station de Biologie des Laurentides of the Université de Montréal. Readers will find the complete description of the sampling site and of the variables in the Borcard *et al.* papers referred to here, as well as some biological information about the lake and its oribatid mite community. The data set is available from the authors. A total of 70 cores, each 5 cm in diameter and 7 cm in depth, was extracted from a 10 × 2.6 m plot in the peat blanket. For the purpose of the analysis, they retained 35 species of mites (matrix **P**) as well as five environmental variables (matrix **E**) (Fig. 2): substratum (seven unordered qualitative classes: four species of *Sphagnum*, ligneous litter, bare peat, interface between *Sphagnum* species), coverage density of the shrub cover (three semi-quantitative classes), microtopography of the substratum (two qualitative classes: blanket and hummock), density of the substratum in $g\,l^{-1}$ of dry uncompressed matter (quantitative variable), and water content in per cent (quantitative variable). These variables will be used in that order in all the following analyses.

The nine terms of a cubic trend surface regression equation have been submitted to the CANOCO procedure of 'forward selection of explanatory variables', and the following five-term equation has been retained:

$$z = b_1x + b_2y + +b_4xy + b_5y^2 + b_9y^3$$

These same terms are used as in matrix **S** of the present paper.

3.2 *Results part I: an obvious gradient structure*

All the analyses – correspondence analysis (CA) of **P**, CCA of **P** constrained by both **E** and **S** (table $\mathbf{P_T}$), CCA of **P** constrained by **E** (table $\mathbf{P_E}$), analysis of the triplet $(\mathbf{P_{EonS}}, \mathbf{D}_i, \mathbf{D}_j)$, CCA of **P** constrained by $\mathbf{O_S}(\mathbf{E})$ (table $\mathbf{P_{O_S(E)}}$) – constrained or not by the environmental variables without removing the space effect, give approximately the same results (Figs 3 and 4 and Table 1); they display a dominant structure reflecting a south-to-north gradient. Note that there does not exist any linear combination of the environmental variables which can be perfectly written using the spatial variables, or vice versa. Hence, the table of predictors $\mathbf{E} \cap \mathbf{S}$ is empty and no analysis can be carried out with this criterion.

Correlations between the environmental variables and the environmental site scores Fig. 4 show that this gradient is mainly explained by humidity. A clear link also appears between most of the other environmental variables and humidity. Shrub cover is close to absent when humidity is large, and abundant in drier areas. Large humidity concerns mainly the flat microtopography while the raised zones (hummock) are drier.

Three kinds of variables appear, however, not to be related to this gradient (Figs 2 and 4): *Sphagnum* distributions, which form regionalized areas not corresponding alone with

Table 1. Correlations among 'row scores' of the various analyses, for the first two canonical eigenvalues. Correlations are calculated for the different analyses without removing the effect of the spatial data table. CA is the direct CA of **P**; PT is the CCA of **P** constrained by both the spatial and environmental factors; PE is the CCA of **P** constrained by the environmental table; PEonS is the analysis of the triplet $(\mathbf{P}_{\mathbf{EonS}}, \mathbf{D}_i, \mathbf{D}_j)$; Os(E) is the CCA of **P** constrained by the $\mathbf{O_s}(\mathbf{E})$ type predictors (see text). For the initial CA, there is only one type of row scores which is used for the calculus of correlations for both environmental and species row score issued from the other analyses.

Axis 1

	CA	PT	PE	PEonS	Os(E)
CA	1	0.98	0.92	0.95	–0.91
PT		1	0.94	0.97	–0.93
PE			1	0.89	–0.99
PEonS				1	–0.90
Os(E)					1

Axis 2

	CA	PT	PE	PEonS	Os(E)
CA	1	0.82	–0.67	–0.66	–0.59
PT		1	–0.80	–0.83	–0.73
PE			1	0.63	0.88
PEonS				1	0.74
Os(E)					1

Environment

Axis 1

	CA	PT	PE	PEonS	Os(E)
CA	1	0.99	0.99	0.99	–0.99
PT		1	0.99	0.99	–0.99
PE			1	0.99	–0.99
PEonS				1	–0.99
Os(E)					1

Axis 2

	CA	PT	PE	PEonS	Os(E)
CA	1	0.99	–0.98	–0.94	–0.96
PT		1	–0.98	–0.97	–0.98
PE			1	0.95	0.98
PEonS				1	0.99
Os(E)					1

Species

parts of the gradient; shrub cover 'present', which shows up in both the positive and negative parts of this principal structure; and density of the substratum, which is certainly the most heterogeneous variable from site to site. Species scores along this axis show that most of their distributions are extremely dependent on the humidity gradient.

The second axis of all analyses gives the plot of the first two axes a horseshoe shape. Several environmental and spatial variables contribute in an important way to this axis. Two kinds of species are, however, well described by this axis: those that show only one small area of presence in the south or in the north and absence everywhere else, and those whose presence is restricted to a northeast–southwest oriented (negative coordinates) area in the centre of the map.

The most surprising result comes from the CCA of **P** constrained by the polynomial after removing the effect of the environment variables (fraction (c) of Borcard *et al.*). The first two axes (Fig. 5) still display large-scale structures oriented south-to-north for the first one, and southwest–northeast for the second. If the species row scores are completely similar to those from the previous analyses (correlations of 0.97 and 0.87 between the first two axes of this analysis and those of the analysis constrained by both the environmental and spatial variables), local or regionalized spatial changes appear much larger for the 'spatial' site scores, which are the most interesting in this study. Since the species scores are scaled in such a way that they come out as the weighted averages of the environmental site scores, some species are shown, under this model, to

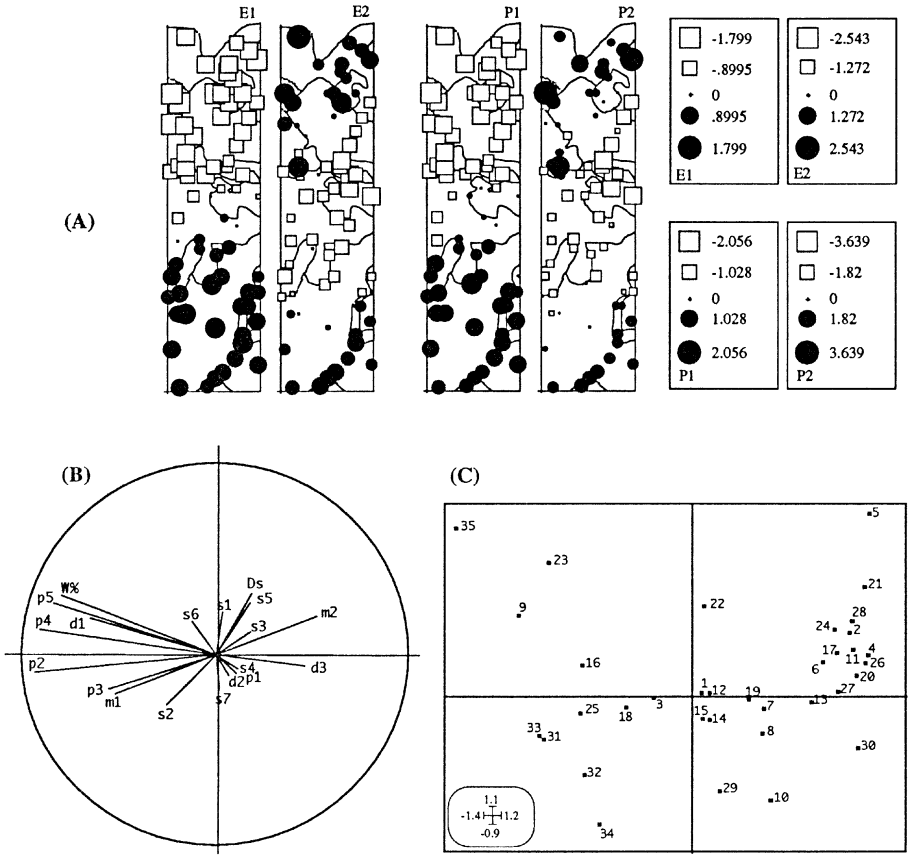

Figure 4. CCA (canonical axes 1 and 2) of the species table constrained by both the environmental and spatial variables. This analysis is presented as a compromise among the analyses without removing the space effect because the absolute values of the correlations between the sampling site scores of this analysis and the others are, on the average, the largest (Table 1). (A) Left: environmental site scores (E); right: species row scores (P). Circles of proportional areas illustrate positive values, while squares represent negative values. The contours of the *Sphagnum* are shown on the maps. (B) Circle of correlations between environmental and spatial variables and environmental site scores: s1 to s7: substratum; m1, m2: microtopography of the substratum; d1 to d3: coverage density of the schrub cover; Ds: density of the substratum; W%: percentage water content; p1 to p5: successive terms of the polynomial. (C) Species scores: species are represented by their positions in the species data table.

be specialists of some small areas, and (or) present great peaks of presence in specific zones. These areas or peaks fit in with larger-scale structures which are globally similar to the ones given by the previous analyses. The difference is that these larger-scale structures do not present continuous variation, but global oppositions with small-scale heterogeneity within the opposite areas.

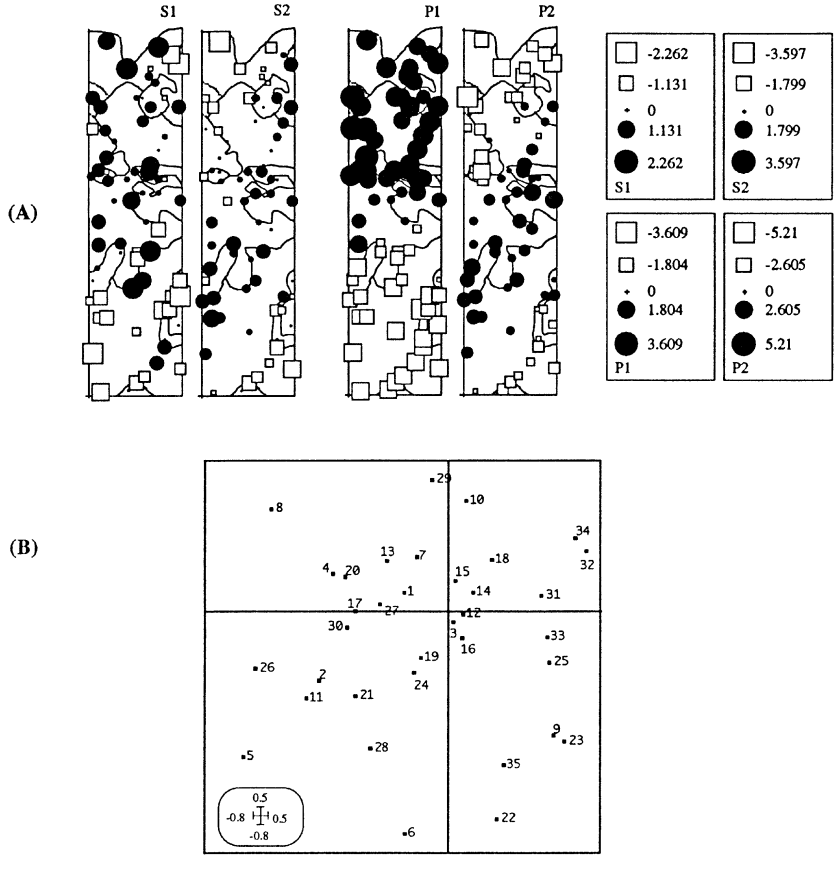

Figure 5. CCA (canonical axes 1 and 2) of the species table constrained by the spatial variables after removing the effect of the environmental predictors. (A) Left: 'environmental' site scores (S). Right: Species row scores (P). (B) Species scores. The environmental variable scores do not figure here because their effect has been removed. Correlations between spatial variables and 'environmental' site scores are not shown because the monomials (i.e. the five terms of the polynomial) do not have a clear interpretation.

3.3 *Results part II: removing the effect of space*

Figure 6 shows the analysis of the table of the residuals of the species table constrained by both the environmental and species tables (table **R**; analysis of the triplet $(\mathbf{R}, \mathbf{D}_i, \mathbf{D}_j)$). The decrease of eigenvalues (Fig. 3) is very small between the first two and the others: no important multivariate structures remain in these residuals. As was foreseeable, the non-environmental site scores are spatially very heterogeneous. Even with this great heterogeneity, a few species appear to be closely linked to the first site scores. These species are mainly characterized by very small isolated patches or peaks of presence or absence. Numbers of individuals of these species in the patches are often very high. Large distances

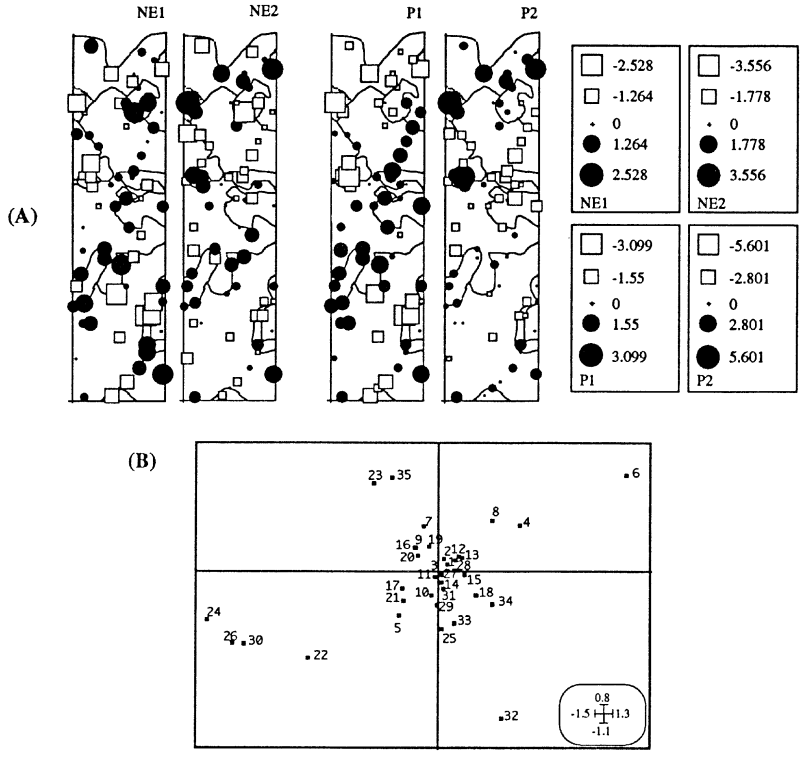

Figure 6. Residual axes of variation (matrix **R**) after analysis of the species data table by the environmental and spatial variables. (A) Left: non-environmental site scores (NE). Right: Species row scores (P). (B) Species scores.

generally separate these zones. Because we have removed the effects of environment and space, these distributions are largely independent of the environmental and spatial variables taken into account; they describe local phenomena.

Figure 7 shows the two analyses that we proposed to study the effects of local environmental predictors on the species distributions. The distributions of eigenvalue **s** (Fig. 3) are quite different: while one eigenvalue is largely dominant in the analysis (ANA1) of $\mathbf{P}_{\mathbf{Eon/S}}$ (residuals from the regression of $\mathbf{P}_{\mathbf{E}}$ on **S**), at least two of them are quite strong in that of $\mathbf{P}_{\mathbf{E}\cap\mathbf{S}^{\perp}}$ (ANA2 on the model of $\mathbf{P}^{\circ}$ by the linear combinations of environmental variables which are independent of any spatial variable). The distribution of eigenvalues in the first case is more discontinuous than in the second case. Visual inspection of the correlations between the environmental variables and the environmental site scores also shows great differences between the two analyses.

Surprisingly, the first axis of ANA1 is essentially linked with humidity. It clearly appears that the areas around the bare peat and flooded areas play an important role in this link because they are among the most humid. Several species show a medium-strength link with this axis, some of them completely avoiding the very humid areas as

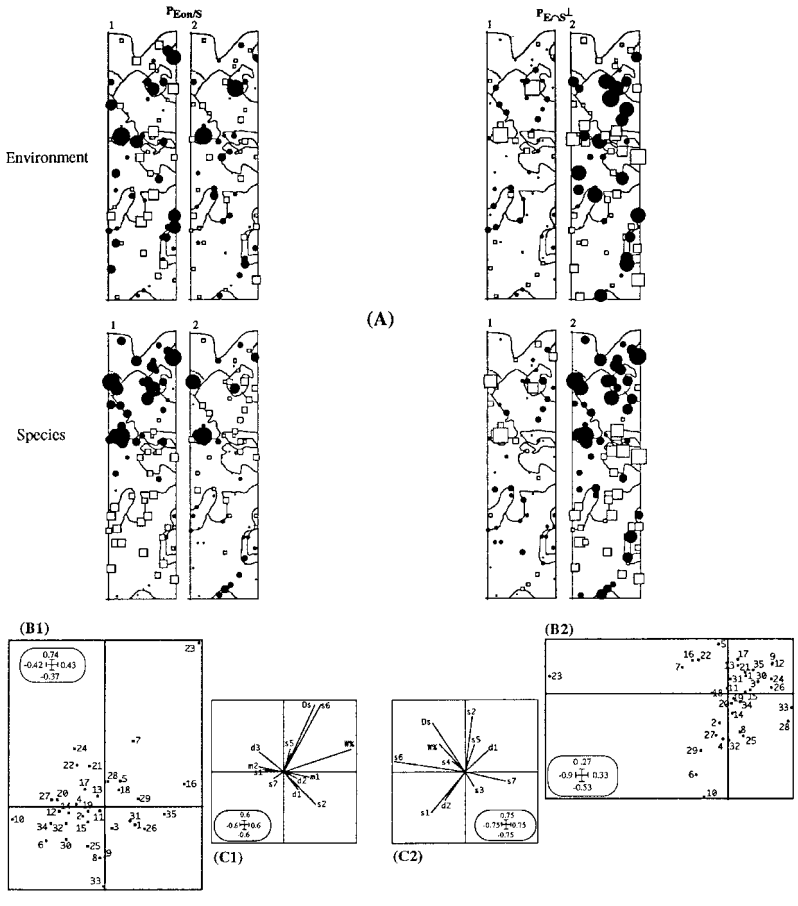

Figure 7. Left: analysis of the triplet $(\mathbf{P}_{\mathbf{Eon/S}}, \mathbf{D}_i, \mathbf{D}_j)$ (model of the species table by the environmental variables, removing the effect of space). Right: CCA of the species table constrained by the linear combinations of the environmental variables, independent of the spatial variables (table $\mathbf{P}_{\mathbf{E}\cap\mathbf{S}^\perp}$). (A) Environment: environmental site scores, axes 1 and 2. Species: species row scores, axes 1 and 2. (B1) and (B2) Species scores in each analysis. (C1) and (C2) Correlations between the environmental variables (same notation as in Fig. 4) and the environmental site scores; for clarity, variables with too small correlations (near zero) are not shown.

well as bare peat. Few of the others seem to have some real preference for these zones. The second axis of this analysis is dependent on the density of the substratum and on bare peat. Only one species shows great affinity with the distribution shown by this axis. It is clearly linked both with bare peat and high density of the substratum.

On the other hand, the first axis of ANA2 is largely determined by bare peat and at a lower level by the borders between substratum classes. One species (*Trhypochthonius* sp., no. 23) dominates this axis; it is the same species as the one that dominates the second axis of ANA1. The second axis mostly depends on *Sphagnum* groups 1 and 2,

and also partially on the density of the substratum. Two species, appearing to be essentially restricted to *Sphagnum* group 2, are particularly important on this axis. Axis 3 (not shown) is linked with *Sphagnum* groups 2 and 3, as well as with the borders between substratum classes. If no species really appears to be strictly linked to these different variables, it is interesting to note that the site scores associated with the environmental table show very homogeneous values for them. It is about the same with the second axis, where sampling sites associated with groups 1 and 2 always have opposite values. This last analysis brings out the effect of the substratum classes, which was not well taken into account in the other approaches.

3.4 *Results part III: discussion*

A quick assessment of the previous analyses shows the following:

- When no constraint of independence to space is imposed, what comes out of the analysis using environmental variables is the humidity gradient, which is very important in the data set under study.
- The residuals from the model of $\mathbf{P}^{\circ}$ by the additive effect of environment and space shows the influence of species distributed in patches or peaks, which are independent from the environmental and spatial variables under consideration. Few species display this kind of distribution.
- The analysis of the species table constrained by the spatial table after removing the effect of the environmental variables shows patchy distributions fitting in with larger-scale structures, which globally correspond to the dominant gradient structure, together with a lot of variation at smaller scale. Very heterogeneous values are present in the differentiated zones.
- The residuals from the model of $\mathbf{P}_{\mathbf{E}}$ by the spatial variables show patchy distributions, linked first with the environmental variable which accounts for the largest amount of the variability of the species table (humidity), and secondly with some other very specific environmental variables (bare peat and density of the substratum). It thus appears that this two-step analysis (model of $\mathbf{P}^{\circ}$ by $\mathbf{E}$, followed by model of ($\mathbf{P}^{\circ}$ by $\mathbf{E}$) by $\mathbf{S}$, from which the residuals are analysed) strongly shows the mark of the first model by putting in evidence the local effects included in larger-scale structures. Because in the first step humidity is the variable which accounts for most of the variability, and presents at the same time some regionalized structures, it remains an important variable in the analysis.
- The model of $\mathbf{P}^{\circ}$ by $\mathbf{E} \cap \mathbf{S}^{\perp}$ is more strongly linked with the strictly regionalized variability, associated with the substratum variables. The correspondence between the environmental site scores and the *Sphagnum* groups is excellent, at least with the most important groups in terms of number of samples (groups 1 and 2) and in terms of local effects (bare peat). It is also true for the two following axes which all translate the effect of the strata. The variability explained by $\mathbf{E} \cap \mathbf{S}^{\perp}$ is very small compared to the total variability, because the substratum is described by a qualitative variable with a lot of modalities which, together, do not play a large role in explaining the variation. It is, however, clear that the strata are regionalized, and practically independent of

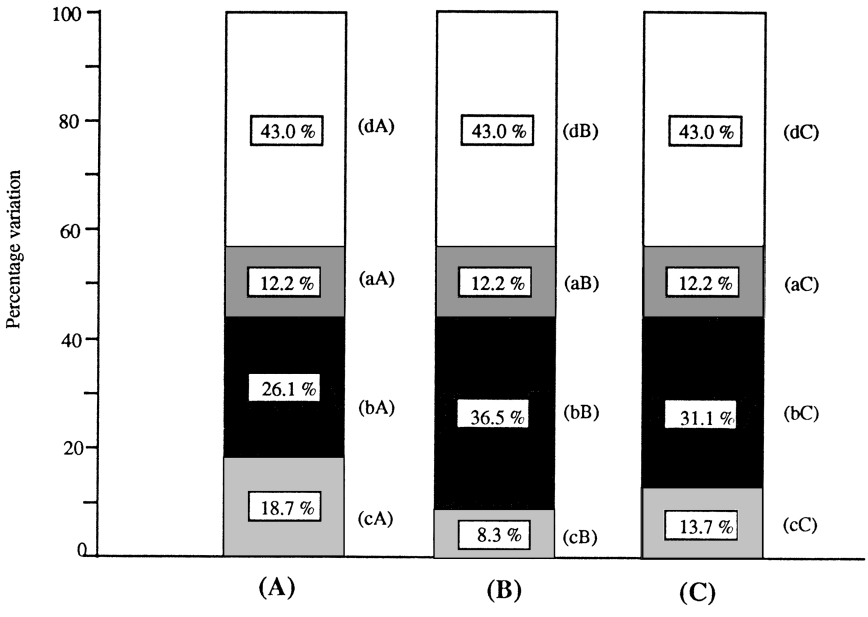

Figure 8. Variation partitioning of the oribatid mite data table. (A) Partition using the regression model of $\mathbf{P_E}$ on the spatial variables (variabilities associated with the model and the residuals). (B) Partition using the decomposition of **E** in one part independent from all the spatial variables and one part neither collinear nor independent from these same variables. (C) Partition of Borcard *et al.*; dA, dB and dC are the percentages of variability associated with the table **R** (residuals from the additive model). aA, aB and aC are the percentages associated with the model $\mathbf{P_{S/E}}$ (model of **P** by spatial predictors from which the effects of the environmental variables have been removed). bA and cA are the percentages associated with the table $\mathbf{P_{EonS}}$ and $\mathbf{P_{Eon/S}}$. bB and cB are the percentages associated with $\mathbf{P_{Os(E)}}$ and $\mathbf{P_{E\cap S^{\perp}}}$. bC and cC are the percentages given by the Borcard *et al.* procedure (variabilities associated with $\mathbf{P_{E/S}}$ and differences of the explained variances by $\mathbf{P_E}$ and $\mathbf{P_{E/S}}$). Note that two different models are at the origin of decompositions (A) and (B), and that the part (bC) of the third does not have a clear meaning. Hence the comparisons could be made by reference to these models. Notice also that the parts linked with the undetermined and 'pure spatial' fractions are always the same and that (cB) and (cC) are directly comparable (see text).

the gradient structure. This part of the variability is thus important to understand. Few oribatid species are distributed in patches dependent upon particular strata, but it is possible to determine them well by this means. The effect of the main environmental variables that are weakly linked with the gradient of humidity is thus well represented.

Although the two possible decompositions do not exactly answer the same questions, a quick summary can be formulated using the same frame of reference as Borcard *et al.* (Fig. 8; see legend for the possible direct comparisons). The first model ((A) in Fig. 8) divides the variability associated with the $\mathbf{P_E}$ model in two not very different parts (bA and cA). As has been said, the explanation can be found in the two-step procedure:

humidity is the strongest explanatory variable, displaying at the same time local and large-scale variations. The second model (B) first shows no variability associated with strictly spatially structured environmental predictors (which actually do not exist in this analysis). Secondly, it exhibits a lot of variability associated with predictors which show at the same time large- and local-scale structures (bB). Finally there is a very small amount of variability associated with local environmental predictors (cB). This variability is however very interesting to consider in order to identify specialized species. This variability is also smaller than that explained by the $\mathbf{P_{E/S}}$ model used by Borcard *et al.* (cC); this is not surprising because the model $\mathbf{P_{E/S}}$ expresses the effects of environmental variables presenting local structures which can be included in larger-scale variation, or not.

There exists a great amount of complementarity between these approaches, when the purpose is to partial out the spatial component. Using the decomposition of $\mathbf{P_E}$ on the space variables gives some explanation about the main explanatory variables (humidity, but not the *Sphagnum* groups), for which effects are also large-scale ($\mathbf{P_{EonS}}$) or regionalized ($\mathbf{P_{Eon/S}}$). Secondly, analysing the model $\mathbf{P_{E/S}}$ (not shown here, but part (a) in Fig. 1) shows both the effects of the environmental variables (mainly the *Sphagnum* groups) which are completely independent of the spatial predictors, and of the environmental variables which display large-scale structures as well as more local ones (like humidity, which presents great regional variations fitting inside the dominant gradient). Finally, analysing the model of $\mathbf{P}$ by $\mathbf{E} \cap \mathbf{S}^{\perp}$ shows the effect of the environmental variables that are completely independent of the spatial predictors (like the *Sphagnum* groups). Note that the two-step procedure corresponding to the analysis of $\mathbf{P_E}$ compared with $\mathbf{S}$ has a specific status in these approaches. Indeed, rather than the effect of spatially structured environmental predictors, it is the patterns of the biotic community which are explained by the environmental predictors, studied with respect to the spatial variables.

4. General discussion

We proposed above two partitioning methods that are complementary to the method of Borcard *et al.* to decompose the variation associated with a species table into different components, determined by environmental and spatial descriptors. Our approach is an extension, to the case of mixed factors, of some results concerning the decomposition of variance in a model with two additive factors and no interaction.

The multivariate analysis includes two steps. First the table to be analysed is decomposed in different parts, using covariables well characterized by their properties of statistical dependence/independence. Secondly, the analyses of the triplets composed by the modelled tables and the scalar products defined by the species table are performed. Depending on the analysed table, these analyses may or may not be equivalent to CCA. This makes it possible to speak about the different effects in terms of models of the species, which allows us to analyse not only the explained variation, but also the multivariate structure of these models by mapping the results of the different analyses, as suggested by Borcard and Legendre (1994). It is particularly true for the equivalent of fraction (b) of Borcard *et al.*, which did not correspond to a model.

Our proposals have tried to respect the purpose of decomposing the variation, as stated by Borcard *et al.* It appears, however, that no single satisfying solution exists to meet their purpose. As in an analysis of variance with unbalanced designs (e.g. Shaw

and Mitchell-Olds, 1993), several approaches have to be tried, which implies favouring some predictors to the detriment of others. Hence it is the choice of the predictors and of their qualities (e.g. explained by spatial variables, completely spatialized, completely independent of any spatial variables, etc.) which is the most important. The extremely well-developed field of linear modelling can allow ecologists to make efficient choices.

We have chosen in this paper to develop the decomposition of the environmental effects by the spatial variables. The symetrical decomposition could also be done to bring out the part of the spatial structure of the community that is independent of the environmental variables, and which is of great interest for ecologists (Borcard and Legendre, 1994); the mathematical framework is strictly the same.

The only decomposition that avoids any confounding is the last one. The constraints are very strong, however. In our illustration for example, no environmental predictor was found to be strictly linked with the spatial variables ($\mathbf{E} \cap \mathbf{S}$). A possible way to reduce this constraint would be to consider, in the canonical correlation analysis of **E** and **S**, all pairs of canonical variables that have large correlations, rather than simply those with unit correlations. For the environmental table, using only these linear combinations as predictors could be a compromise between environmental predictors that are completely or incompletely spatially structured. This approach remains to be explored.

When favouring a decomposition of the environmental effects, no other way is possible to reduce these constraints and obtain at the same time a complete decomposition into linearly independent components. But as we have seen with the oribatid mite data set, this very constrained decomposition is efficient in emphasizing species that are distributed in patches as a response to environmental conditions that present regionalized or peak distributions not included in the main gradient structure.

Leaving aside the partitioning purpose, the two sets of descriptors can be used as the basis for several different models of the species data table (Table 2), which could be compared in terms of explained variability for a given type of predictors. While the interpretation of some of them is interesting, others have no clear biological meaning.

The growing interest among ecologists and environmental scientists for the integration of space into ecological studies can find here efficient procedures aimed at separating large-scale from more regionalized structures, and associating these to possible explanations in terms of environmental variables, historical dynamics, etc. (Borcard and Legendre, 1994, Table 3).

It must be remembered, however, that trend surface analysis presents problems; one of the most obvious is the strong correlations among the monomial terms. These monomials must be considered more as an *ad hoc* way to study or eliminate large-scale structures than as a universal tool for modelling them. Because distributions of species can take several forms in space (from noise to gradients), it is important to consider other more specialized tools for their study. In the same framework of linear modelling as used here, recent proposals using neighbourhood relationships (Méot *et al.*, 1993; Legendre and Borcard, 1994; Thioulouse *et al.*, 1995a) would allow differentiation between large-scale, local, and more regionalized structures. Some models derived from multivariate geostatistics (Grzebyk and Wackernagel, 1994; Bourgault and Marcotte, 1991) can also give interesting conceptual ideas about the different scales of variation and the continuity of multivariate distributions.

Table 2. Different models to explore the effects of particular predictors; 'on' means 'regression on'. $\mathbf{P_S}$ and $\mathbf{P_{/S}}$ are the model and residuals from the regression of the species variables on the spatial variables

Model	Interpretation
Models linked with spatially structured environmental predictors	
P on (E on S)	Effect of environmental variables partly or totally spatialized.
P_E on S	Environmental variation of the biotic community which is spatialized (example in the text).
P on $E \cap S$	Effect of environmental variables totally spatialized.
Models linked with non spatial environmental predictors	
P on E/S	Effect of environmental variables which show regional variations fitting with or not in larger-scale structures ((a) in Fig. 1).
$P_{Eon/S}$	Environmental variation of the biotic community which is not spatialized (text).
P on $E \cap S^{\perp}$	Effect of environmental variables which show only regional structures (text) (stronger equivalent to (a)).
Models including spatial predictors linked with environment	
P on (S on E)	Effect of the spatial variables partly or totally linked with environmental predictors (no clear meaning).
P_S on E	Spatial variation of the biotic community which can be explained by environmental predictors.
P on $S \cap E$	Same as P on $E \cap S$.
Models including spatial predictors not linked with environment	
P on S/E	Effect of spatial variables which are partly or totally not linked with environmental predictors ((c) in Fig. 1).
$P_{Son/E}$	Spatial variation of the biotic community which can not be explained by environmental predictors.
P on $S \cap E^{\perp}$	Effect of spatial variables which are totally unlinked with environmental traits (stronger equivalent to (c)).

Acknowledgements

We wish to thank three anonymous referees, whose helpful comments and suggestions have allowed us to greatly improve the manuscript.

The French 'Ministère de l'Enseignement Supérieur et de la Recherche' provided financial support for the postdoctoral leave. This work has also been supported by NSERC grant No. OGP0007738 to P. Legendre.

References

Afriat, S.N. (1957) Orthogonal and oblique projectors and the characteristics of pairs of vector spaces. *Proceedings of the Cambridge Philosophical Society, Mathematical and Physical Sciences*, **53**, 800–16.

Borcard, D., Legendre, P. and Drapeau, P. (1992) Partialling out the spatial component of ecological variation. *Ecology*, **73**, 1045–55.

Borcard, D. and Legendre, P. (1994) Environmental control and spatial structure in ecological communities: an example using oribatid mites (Acari, Oribatei). *Environmental and Ecological Statistics*, **1**, 37–53.

Bourgault, G. and Marcotte, D. (1991) Multivariable variogram and its application to the linear model of coregionalization. *Mathematical Geology*, **23**, 899–928.

Cailliez, F. and Pagès, J.-P. (1976) *Introduction à l'analyse des données*. SMASH, Paris.

Cliff, A.D., and Ord, J.K. (1981) *Spatial Processes*. Pion, London.

Chessel, D., Lebreton, J.-D. and Yoccoz, N. (1987). Propriétés de l'analyse canonique des correspondances. Une illustration en hydrobiologie. *Revue de Statistique Appliquée*, **35**, 55–72.

Chévenet, F., Dolédec, S. and Chessel, D. (1994) A fuzzy coding approach for the analysis of long-term ecological data. *Freshwater Biology*, **31**, 295–309.

Denslow, J.S. (1985) Disturbance-mediated coexistence of species. In *The Ecology of Natural Disturbance and Patch-dynamics*. S.T.A. Pickett and P.S. White (eds) Academic Press, Orlando, pp. 307–23.

Dolédec, S., Chessel, D., ter Braak, C.J.F. and Champely, S. (1996) Matching species traits to environmental variables: a new three-table ordination method. *Environmental and Ecological Statistics*, **3**, 143–66.

Escoufier, Y. (1982) L'analyse des correspondances des tableaux simples et multiples. *Metron*, **40**, 53–77.

Escoufier, Y. (1987) The duality diagram: a means for better practical applications. In *Development in Numerical Ecology*, P. Legendre and L. Legendre (eds), Springer-Verlag, Berlin, pp. 139–56.

Fraile, L., Escoufier, Y. and Raibaut, A. (1993) Analyse des correspondances de données planifiées: étude de la chémotaxie de la larve infestante d'un parasite. *Biometrics*, **49**, 1142–53.

Graybill, F.A. (1976). *Theory and Application of the Linear Model*. Wadsworth Publishing Co., Belmont.

Greenacre, M.J. (1984) *Theory and Applications of Correspondence Analysis*. Academic Press, London.

Gifi, A. (1990) *Nonlinear Multivariate Analysis*.Wiley, Chichester.

Gittins, R. (1985) *Canonical Analysis, A Review with Applications in Ecology*. Springer-Verlag, Berlin.

Griffith, D.A. (1988) *Advanced Spatial Statistics*. Kluwer, Dordrecht.

Grzebyk, M. and Wackernagel, H. (1994) Multivariate analysis and spatial/temporal scales: real and complex models. In *Proceedings (Invited Papers) of the XVIIth International Biometric Conference*, Hamilton, Canada, pp. 19–33.

Hubert, L.J., Colledge, R.G. and Costanzo, C.M. (1981) Generalized procedures for evaluating spatial autocorrelation. *Geographical Analysis*, **13**(3), 224–33.

Isaaks, E.H. and Srivastava, R.M. (1989) *An Introduction to Applied Geostatistics*. Oxford University Press, New York.

Lebreton, J.D., Sabatier, R., Banco, G. and Bacou, A.M. (1991) Principal component and correspondence analyses with respect to instrumental variables: an overview of their role in studies of structure-activity and species-environment relationships. In *Applied Multivariate Analysis in SAR and Environmental Studies*, J. Devillers and W. Karcher (eds) Kluwer, Dordrecht, pp. 85–114.

Legendre, P. and Borcard, D. (1994) Rejoinder. *Environmental and Ecological Statistics*, **1**, 57–61.

Legendre, P. and Fortin, M.-J. (1989) Spatial pattern and ecological analysis. *Vegetatio*, **80**, 107–38.

Legendre, P. and Troussellier, M. (1988) Aquatic heterotrophic bacteria: Modelling in the presence of spatial autocorrelation. *Limnology and Oceanography*, **33**, 1055–67.

Mantel, N. (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**, 209–20.

May, R.M. (1984) An overview: real and apparent patterns in community structure. In *Ecological Communities: Conceptual Issues and the Evidence*, D.R. Strong, D. Simberloff, L.G. Abele and A.B. Thistle (eds) Princeton University Press, Princeton, NJ, pp. 3–16.

Méot, A., Chessel, D. and Sabatier, R. (1993) Opérateurs de voisinage et analyse des données spatio-temporelles. In *Biométrie et Environnement*, J. D. Lebreton and B. Asselain (eds) Masson, Paris, pp. 45–71.

Pernin, M.-O. and Pagés, M. (1988) Comparaison de deux méthodes: l'analyse des correspondances complète et l'analyse dissymétrique conditionnelle. *Statistique et Analyse des Données*, **13**(3), 44–55.

Pontier, J. and Pernin, M.-O. (1987) Solution using 'LONGI'. In *Data Analysis: Ins and Outs of Solving Real Problems*, J. Jansen, F. Marcotorchino and J.M. Proth (eds). Plenum, New York, pp. 49–66.

Pontier, J., Dufour, A.-B. and Normand, M. (1990) *Le modèle euclidien en analyse des données*. SMA, édition Ellipses, Bruxelles.

Prodon, R. and Lebreton, J.D. (1994) Analyses multivariées des relations espèces-milieu: structure et interprétation écologique. *Vie Milieu*, **44**, 69–91.

Rao, C.R. (1964) The use and interpretation of principal component analysis in applied research. *Sankhyā A*, **26**, 329–58.

Rao, C.R., and Yanaï, H. (1979) General definition and decomposition of projectors and some applications to statistical problems. *Journal of Statistical Planning and Inference*, **3**, 1–15.

Rossi R.E., Mulla, D.J., Journel, A.G. and Franz, E.H. (1992) Geostatistical tools for modelling and interpreting ecological spatial dependence. *Ecological Monographs*, **62**(2), 277–314.

Sabatier, R., Lebreton, J.-D. and Chessel, D. (1989) Principal component analysis with instrumental variables as a tool for modelling composition data. In *Multiway Data Analysis*, R. Coppi and S. Balasco (eds) Elsevier, Amsterdam. pp. 341–52.

Shaw, R.G. and Mitchell-Olds, T. (1993) Anova for unbalanced data: an overview. *Ecology*, **74**, 1045–55.

Smouse, P.E., Long, J.C. and Sokal, R.R. (1986). Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Systematic Zoology*, **35**, 627–32.

Sokal, R.R. and Oden, N.L. (1978) Spatial autocorrelation in biology. 2. Some biological implications and four applications of evolutionary and ecological interset. *Biological Journal of the Linnean Society*, **10**, 229–49.

Takeuchi, K., Yanai, H. and Mukerjee, B.N. (1982) *The Foundations of Multivariate Analysis. A Unified Approach by Means of Projection on Linear Subspaces*. Wiley Eastern, New Delhi.

ter Braak, C.J.F. (1986) Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, **67**, 1167–79.

ter Braak, C.J.F. (1988) Partial canonical correspondence analysis. In: *Classification and Related Methods of Data Analysis*, H.H. Block (ed.) North Holland, Amsterdam, pp. 551–8.

ter Braak, C.J.F. (1990) *Update notes: CANOCO version 3.10*. Agricultural Mathematics Group, Wageningen, The Netherlands.

ter Braak, C.J.F. and Verdonschot, P.F.M. (1995) Canonical correspondence analysis and related multivariate methods in aquatic ecology. *Aquatic Sciences*, **57**(3), 255–89.

Thioulouse, J., Chessel, D. and Champely, S. (1995a) Multivariate analysis of spatial patterns: a unified approach to local and global structures. *Environmental and Ecological Statistics*, **2**, 1–14.

Thioulouse, J., Dolédec, D., Chessel, D. and Olivier, J.M. (1995b) ADE software: multivariate analysis and graphical display of environmental data. In: *Proceedings of the 4th International Software Exhibition for Environmental Science and Engineering*, pp. 57–62.

Upton, G. and Fingleton, B. (1985) *Spatial Data Analysis by Example: Point Pattern and Quantitative Data*. Vol. 1. Wiley, New York.

Whittaker, J. (1984) Model interpretation from the additive elements of the likelihood function. *Applied Statistics*, **33**(1), 52–64.

Yoccoz, N. and Chessel, D. (1988) Ordination sous contraintes de relevés d'avifaune: élimination d'effets dans un plan d'observations à deux facteurs. *Compte rendu hebdomadaire des séances de l'Académie des sciences*, série III, **307**, 189–94.

Appendix 1: analysis of a triplet $(\mathbf{M}, \mathbf{D}_i, \mathbf{D}_j)$

This is the outline of the analysis of a triplet $(\mathbf{M}, \mathbf{D}_i, \mathbf{D}_j)$ where $\mathbf{M}$ is a model of a table $\mathbf{Y}$ by another table $\mathbf{X}$ ($\mathbf{Y_X}$: table of the fitted values; $\mathbf{Y_{/X}}$: table of the residuals), using weighted linear regressions with weights $\{p_{i.}\}$.

Call $\mathbf{M}$ one of the tables $\mathbf{Y_X}$ or $\mathbf{Y_{/X}}$. The analysis of the triplet $(\mathbf{M}, \mathbf{D}_i, \mathbf{D}_j.)$ comprises the following elements:

1. The total inertia is given by the trace of matrix $\mathbf{S} = \mathbf{M^t D_i M D_j}$.
2. The eigenvectors of matrix $\mathbf{S}$, noted $\mathbf{u}$, make up a reference orthonormal basis of the spaces where rows are represented. The ratio of a canonical eigenvalue (denoted by λ) on the inertia of $\mathbf{M}$ gives the percentage of variability (inertia) in that table which is explained by the corresponding canonical axis. The ratio of an eigenvalue on the inertia of table $\mathbf{Y}$ gives the percentage of the inertia of $\mathbf{Y}$ explained by this axis.
3. The row scores resulting from this analysis are standardized by the row margins of $\mathbf{P}$, using

   $$\mathbf{R_e} = \mathbf{M D}_j u / \sqrt{\lambda}.$$

   When $\mathbf{Y} = \mathbf{P}^\circ$, they are called 'Sample scores which are linear combinations of environmental variables' in the output of the CANOCO program. By analogy, we call them 'environmental site scores' in the present paper, whatever the nature of the matrix $\mathbf{Y}$ to be analysed. In some cases, matrix $\mathbf{Y}$ is not matrix $\mathbf{P}^\circ$. For $\mathbf{M} = \mathbf{Y_{/X}}$, we call them 'non-environmental site scores' (independence of the $\mathbf{X}$-variables).
4. The equivalent to 'species scores' of CCA are given by the relationship

   $$\mathbf{S_s} = \mathbf{M^t D}_i \mathbf{R_e} = \mathbf{Y^t D}_i R_e.$$

   We keep this denomination both in the strict CCA case ($\mathbf{Y} = \mathbf{P}^\circ$) or in the more general one ($\mathbf{Y} \neq \mathbf{P}^\circ$).
5. The equivalent to 'sample scores in species space' of CCA are calculated as

   $$\mathbf{R_s} = \mathbf{Y D}_j \mathbf{S_s} / \lambda.$$

   We call them 'species row scores' whatever the nature of matrix $\mathbf{Y}$ to be analysed.
6. There are several alternative ways of building environmental variable scores (regression/canonical coefficients of the environmental variables, centroids of qualitative predictors, etc.). In this paper, we use the correlations (using weighting $\{p_{i.}\}$) between the environmental variables and the environmental site scores.

Appendix 2: Calculation of canonical coefficients in canonical correlation analysis

The canonical coefficients for the variables of tables **E** and **S** are the components of the eigenvectors associated with the two matrices:

$$\mathbf{A}_{EE} = (\mathbf{E}^t\mathbf{D}_i\mathbf{E})^{-1}\mathbf{E}^t\mathbf{D}_i\mathbf{S}(\mathbf{S}^t\mathbf{D}_i\mathbf{S})^{-1}\mathbf{S}^t\mathbf{D}_i\mathbf{E}.$$

$$\mathbf{A}_{SS} = (\mathbf{S}^t\mathbf{D}_i\mathbf{S})^{-1}\mathbf{S}^t\mathbf{D}_i\mathbf{E}(\mathbf{E}^t\mathbf{D}_i\mathbf{E})^{-1}\mathbf{E}^t\mathbf{D}_i\mathbf{S}.$$

Note that generalized inverses (e.g. Graybill, 1976, for their use in linear modelling) are used when there exist linear dependency problems among variables of one of the two sets. The eigenvalues of $\mathbf{A}_{EE}$ and $\mathbf{A}_{SS}$ are positive or null. The strictly positive eigenvalues, denoted by λ, are the same for $\mathbf{A}_{EE}$ and $\mathbf{A}_{SS}$ and they have the same number of associated eigenvectors. They have been normalized in such a way that $\mathbf{u}^t(\mathbf{E}^t\mathbf{D}_i\mathbf{E})\mathbf{u} = 1$ and $\mathbf{v}^t(\mathbf{S}^t\mathbf{D}_i\mathbf{S})\mathbf{v} = 1$, which ensures that the canonical variables have a variance of 1. Eigenvectors associated with strictly positive eigenvalues, noted **u** for $\mathbf{A}_{EE}$ and **v** for $\mathbf{A}_{SS}$, are linked by the relationships:

$$\mathbf{v} = (1/\sqrt{\lambda})(\mathbf{S}^t\mathbf{D}_i\mathbf{S})^{-1}\mathbf{S}^t\mathbf{D}_i\mathbf{E}\mathbf{u}.$$

$$\mathbf{u} = (1/\sqrt{\lambda})(\mathbf{E}^t\mathbf{D}_i\mathbf{E})^{-1}\mathbf{E}^t\mathbf{D}_i\mathbf{S}\mathbf{v}.$$

Biographaphical sketches

Alain Méot was a postdoctoral fellow in the laboratory of Professor P. Legendre from August 1994 to April 1995. His main interest is the introduction of spatial information in ordination techniques. His Ph.D. research, conducted in the Laboratoire de Biométrie, Génétique et Biologie des Populations of Université Claude Bernard-Lyon I in France, dealt with the use of geographic or temporal neighbourhood relationships in ordination techniques, and applications of these methods in agronomy, particularly to describe the functioning of agrosytems. At present, he teaches statistics and data analysis at the Université Clermont-Ferrand II.

Pierre Legendre is professor of quantitative biology at the Université de Montréal. Fellow of the Royal Society of Canada and former Killam Research Fellow (1989–91), he received in 1994 the Distinguished Statistical Ecologist Award of the International Congress of Ecology (INTECOL) and in 1995 the Romanowski Medal (environmental science) of the Royal Society of Canada. He is the author of over 125 refereed articles, and over 250 papers presented at scientific meetings and research seminars, dealing with numerical ecology, community ecology, environmental assessment, spatial analysis and phylogenetics, as well as textbooks (in French and English) on numerical ecology.

Daniel Borcard is a community ecologist interested in soil and peat-bog fauna. After several years spent at Université de Neuchâtel, Switzerland, doing research on soil ecology, he is presently working with Pierre Legendre at Universìté de Montréal on the analysis of spatial and multiscale structures in ecological data.