Analyse canonique,
partition de la variation et analyse CPMV

Legendre, P. 2005. Analyse canonique, partition de la variation et analyse CPMV. *Sémin-R*, atelier conjoint GREFi-CRBF d'initiation au langage R. Université du Québec à Montréal (UQAM), 18 avril 2005.

J'expliquerai d'abord les bases conceptuelles et algébriques de l'analyse canonique de redondance (ACR) et de l'analyse canonique des correspondances (ACC), y compris les transformations pour tableaux d'abondances d'espèces avant l'analyse de redondance (fonction 'transfodata' en langage R). Puis je montrerai comment utiliser R pour obtenir une partition de la variation d'un tableau-réponse d'abondances d'espèces en fonction de tableaux explicatifs de variables environnementales et spatiales. Dans ce type d'analyse, les relations spatiales peuvent être représentées soit à l'aide d'un polynôme algébrique des coordonnées géographiques des sites, soit par des fonctions de base CPMV (fonction 'PCNM' en R). Celles-ci représentent la variation à toutes les échelles spatiales qui peuvent être étudiées par un plan d'échantillonnage.

Plan de l'exposé

1. Analyse canonique de redondance, ACR (*RDA*)

Pages extraites du chapitre 11 de

Legendre, P. and L. Legendre. 1998. *Numerical ecology, 2nd English edition*. Elsevier Science BV, Amsterdam. xv + 853 pages.

2. Analyse canonique des correspondances, ACC (*CCA*)

3. ACR et ACC partielles : voir le fichier PowerPoint (section 6) et la fonction R (section 7)

4. Tests de signification en analyse canonique

5. Manova réalisée par RDA

Une page expliquant le codage d'un facteur par des variables binaires. Le codage de deux ou plusieurs facteurs, ainsi que leur interaction, est décrit à l'annexe C de l'article suivant:

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

6. Partition de la variation, analyse CPMV: fichier PowerPoint. Diapo 37: Références.

7. Fonctions 'rdaTest' et 'graph.rdaTest': voir "rdaTest" dans le dossier des travaux pratiques.

8. Fonction 'PCNM'

Distribution de programmes et de tirés à part en PDF

http://www.bio.umontreal.ca/legendre/

Chapter

11 *Canonical analysis*

11.0 Principles of canonical analysis

Canonical analysis is the simultaneous analysis of two, or eventually several data tables. It allows ecologists to perform a *direct comparison* of two data matrices ("direct gradient analysis"; Fig. 10.4, Table 10.1). Typically, one may be interested in the relationship between a first table describing species composition and a second table of environmental descriptors, observed *at the same locations*; or, say, a table about the chemistry of lakes and another about drainage basin geomorphology.

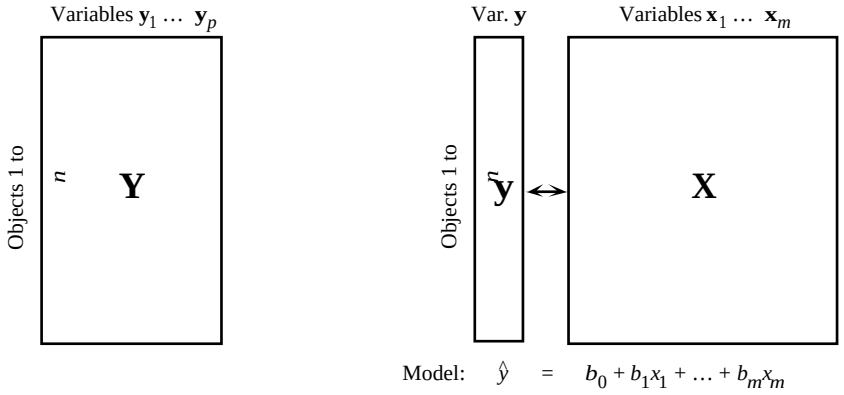
In *indirect comparison* (indirect gradient analysis; Section 10.2, Fig. 10.4), the matrix of explanatory variables **X** does not intervene in the calculation producing the ordination of **Y**. Correlation or regression of the ordination vectors on **X** are computed *a posteriori*. In *direct comparison analysis*, on the contrary, matrix **X** intervenes in the calculation, forcing the ordination vectors to be maximally related to combinations of the variables in **X**. This description applies to all forms of canonical analysis and in particular to the asymmetric forms described in Sections 11.1 to 11.3. There is a parallel in cluster analysis, when clustering results are constrained to be consistent with temporal (Subsection 12.6.4) or spatial relationships (Subsection 13.3.2) among observations, which are inherent to the sampling design. When using a constraint (clustering, ordination), the results should differ from those of unconstrained analysis and be, hopefully, more readily interpretable. Thus, direct comparison analysis allows one to directly test *a priori* ecological hypotheses by (1) bringing out *all* the variance of **Y** that is related to **X** and (2) allowing formal tests of these hypotheses to be performed, as detailed below. Further examination of the unexplained variability may help generate new hypotheses, to be tested using new field observations (Section 13.5).

Canonical form

In mathematics, a *canonical form* (from the Greek κανων, pronounced "kanôn", rule) is the simplest and most comprehensive form to which certain functions, relations, or expressions can be reduced without loss of generality. For example, the canonical form of a covariance matrix is its matrix of eigenvalues. In general, methods of canonical analysis use eigenanalysis (i.e. calculation of eigenvalues and eigenvectors), although some extensions of canonical analysis have been described that use multidimensional scaling (MDS) algorithms (Section 9.3).

(a) Simple ordination of matrix **Y**:
principal comp. analysis (PCA)
correspondence analysis (CA)

(b) Ordination of **y** (single axis) under constraint of **X**: multiple regression



(c) Ordination of **Y** under constraint of **X**:
redundancy analysis (RDA)
canonical correspondence analysis (CCA)

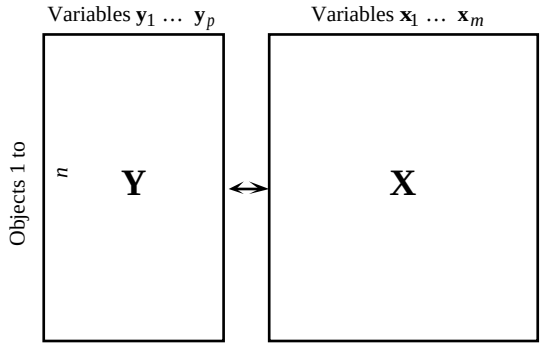


Figure 11.1 Relationships between (a) ordination, (b) regression, and (c) the asymmetric forms of canonical analysis (RDA and CCA). In (c), each canonical axis of **Y** is constrained to be a linear combination of the explanatory variables **X**.

Canonical analysis combines the concepts of ordination and regression. It involves a response matrix **Y** and an explanatory matrix **X** (names used throughout this chapter). Like the other ordination methods (Chapter 9; Fig. 11.1a), canonical analysis produces (usually) orthogonal axes from which scatter diagrams may be plotted.

Canonical analysis has become an instrument of choice for ecological analysis. A 1994 bibliography of ecological papers on the subject already contained 379 titles (Birks *et al.*, 1994). CCorA and discriminant analysis are readily available in most major statistical packages. For RDA and CCA, one must rely on specialized ordination packages. The most widely used program is CANOCO* (ter Braak, 1988b). A closely related procedure, called ACPVI (principal component analysis with instrumental variables), is available in the ADE-4 package† (Thioulouse *et al.*, 1996).

11.1 Redundancy analysis (RDA)

Redundancy

Redundancy analysis (RDA) is the direct extension of multiple regression to the modelling of multivariate response data. *Redundancy* is synonymous with *explained variance* (Gittins, 1985). The analysis is asymmetric: **Y** is the table of response variables and **X** is the table of explanatory variables. Looking at the matter from a descriptive perspective, one would say that the ordination of **Y** is constrained in such a way that the resulting ordination vectors are linear combinations of the variables in **X**. The difference between RDA and canonical correlation analysis (CCorA, Section 11.4) is the same as that between simple linear regression and linear correlation analysis. RDA may also be seen as an extension of principal component analysis (Section 9.1), because the canonical ordination vectors are linear combinations of the response variables **Y**. This means that each ordination vector is a one-dimensional projection of the distribution of the objects in a space that preserves the Euclidean distances (D_1, Chapter 7) among them. These ordination vectors differ, of course, from the principal components that could be computed on the **Y** data table, because they are also constrained to be linear combinations of the variables in **X**.

Redundancy analysis was first proposed by Rao (1964); in his 1973 book (p. 594-595), he proposed the problem as an exercise at the end of his Chapter 8 on multivariate analysis. The method was later rediscovered by Wollenberg (1977).

The eigenanalysis equation for redundancy analysis

$$(\mathbf{S_{YX}}\,\mathbf{S}^{-1}_{\mathbf{XX}}\,\mathbf{S'_{YX}} - \lambda_k\mathbf{I})\,\mathbf{u}_k = \mathbf{0} \qquad \textbf{(11.3)}$$

---

* CANOCO, which contains procedures for both RDA and CCA, was written by C. J. F. ter Braak who also developed CCA. Distribution: see Table 13.4, p. 784.

The package PC-ORD contains a procedure for CCA. Distribution: see footnote in Section 9.3.

RDACCA is a FORTRAN program for RDA and CCA written by P. Legendre. It is distributed free of charge from the WWWeb site: <http://www.fas.umontreal.ca/BIOL/legendre/>. It uses the direct eigenanalysis methods described in Subsections 11.1.1(for RDA) and 11.2.1 (for CCA).

† The ADE-4 package (for Macintosh and Windows) was written by D. Chessel and J. Thioulouse at Université de Lyon, France. It is distributed free of charge from the following WWWeb site: <http://biomserv.univ–lyon1.fr/ADE–4.html>.

may be derived through multiple linear regression, followed by principal component decomposition (Fig. 11.2). This way of looking at the calculations makes it intuitively easy to understand what RDA actually does to the data. It also shows that the computations can be carried out using any standard general statistical package for micro-computer or mainframe, provided that multiple regression and principal component analysis are available; the procedure is also easy to program using advanced languages such as MATLAB or S-PLUS. RDA is appropriate when the response data table **Y** could be analysed, alone, by principal component analysis (PCA); in other words, when the **y** variables are linearly related to one another and the Euclidean distance is deemed appropriate to describe the relationships among objects in factorial space. The data matrices must be prepared as follows, prior to RDA.

1. The table of response variables **Y** is of size ($n \times p$), where n is the number of objects and p is the number of variables. Centre the response variables on their means, or standardize them by column if the variables are not dimensionally homogeneous (e.g. a mixture of temperatures, concentrations, pH values, etc.), as one would do prior to PCA. Centring at this early step simplifies several of the equations from 11.4 to 11.12 in which, otherwise, the centring of the columns of matrix **Y** should be specified.

2. Table **X** of the explanatory variables is of size ($n \times m$) with $m \leq n$. The variables are centred on their respective means for convenience; centring the variables in **X** and **Y** has the effect of eliminating the regression intercepts, thus simplifying the interpretation without loss of pertinent information. The **X** variables may also be standardized (eq. 1.12). This is not a necessary condition for a valid redundancy analysis, but removing the scale effects of the physical dimensions of the explanatory variables (Subsection 1.5.4) turns the regression coefficients into standard regression coefficients which are comparable to one another. The amount of explained variation, as well as the fitted values of the regression, remain unchanged by centring or standardization of the variables in **X**. In the program CANOCO, for instance, standardization is automatically performed for the explanatory variables (matrix **X**) when computing RDA or CCA.

The distributions of the variables should be examined at this stage, as well as bivariate plots within and between the sets **Y** and **X**. Transformations (Section 1.5) should be applied as needed to linearize the relationships and make the distributions more symmetric, reducing the effect of outliers.

If **X** and **Y** are made to contain the same data (i.e. **X** = **Y**), eq. 11.3 becomes $(\mathbf{S_{YY}} - \lambda_k \mathbf{I})\ \mathbf{u}_k = \mathbf{0}$, which is the equation for principal component analysis (eq. 9.1). The result of RDA is then a principal component analysis of that data table.

1 — The algebra of redundancy analysis

The following algebraic development describes how to arrive at eq. 11.3 through multiple regression and principal component analysis. The steps are (Fig. 11.2): (1) regress each variable in **Y** on all variables in **X** and compute the fitted values;

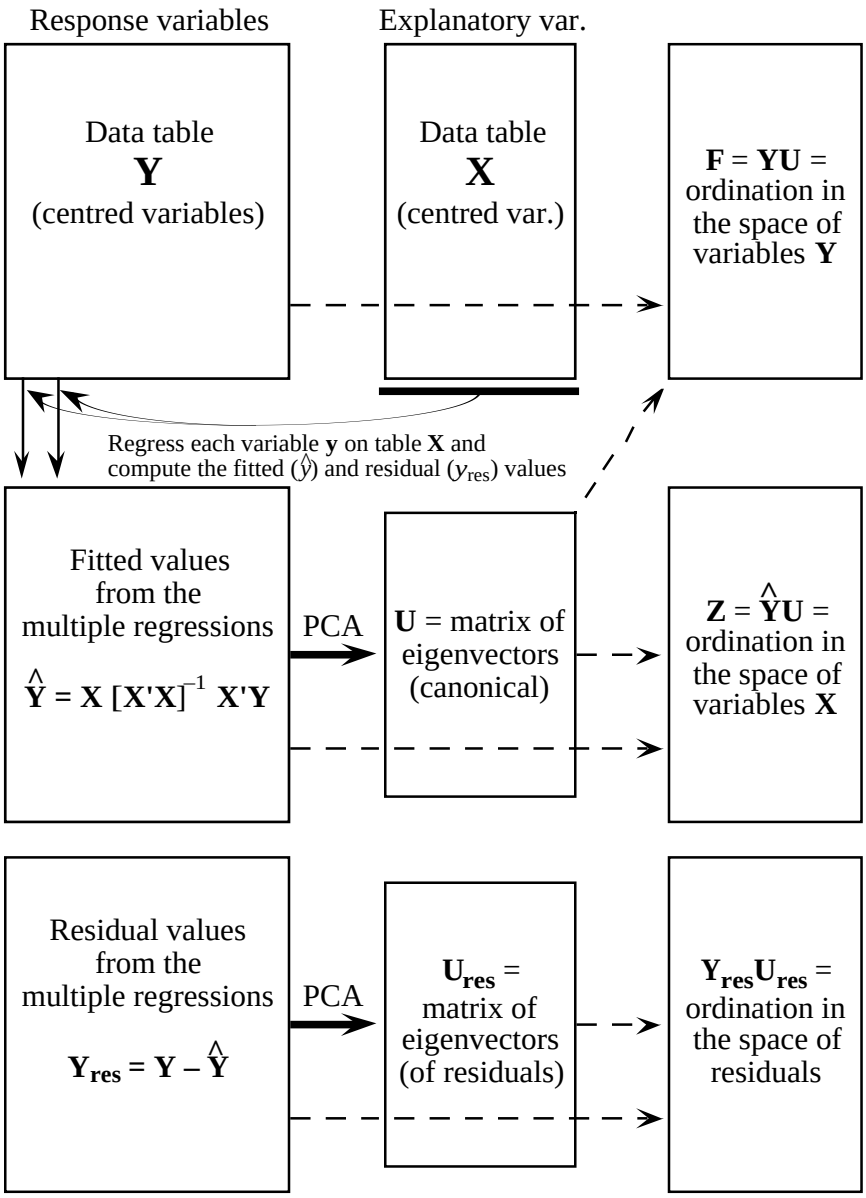


Figure 11.2 Redundancy analysis may be understood as a two-step process: (1) regress each variable in **Y** on all variables in **X** and compute the fitted values; (2) carry out a principal component analysis of the matrix of fitted values to obtain the eigenvalues and eigenvectors. Two ordinations are obtained, one (**YU**) in the space of the response variables **Y**, the other ($\hat{\mathbf{Y}}\mathbf{U}$) in the space of the explanatory variables **X**. Another PCA ordination can be obtained using the matrix of residuals.

(2) carry out a principal component analysis on the matrix of fitted values to obtain the eigenvalues and eigenvectors.

1) For *each* response variable in table **Y**, compute a multiple linear regression on all variables in table **X**. This may be done using any general-purpose statistical package. The matrix equation corresponding to each regression is (eq. 2.19):

$$\mathbf{b} = [\mathbf{X}'\mathbf{X}]^{-1}\,\mathbf{X}'\mathbf{y}$$

so that the matrix equation corresponding to the whole set of regressions (i.e. for all response variables) is

$$\mathbf{B} = [\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Y} \qquad \textbf{(11.4)}$$

where **B** is the matrix of regression coefficients of all response variables **Y** on the regressors **X**. Computing all linear regressions simultaneously has been called *multivariate linear regression* by Finn (1974) and is available, for instance, in the SAS procedure GLM.

In multiple regression, the fitted values $\hat{y}$ are computed as:

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{B} \qquad \textbf{(11.5)}$$

This is the multivariate extension of eq. 10.1. The whole table of fitted values, $\hat{\mathbf{Y}}$, may be computed in a single matrix operation in this way. Using **B** estimated by eq. 11.4, eq. 11.5 becomes:

$$\hat{\mathbf{Y}} = \mathbf{X}[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Y} \qquad \textbf{(11.6)}$$

Because variables **X** and **Y** are centred on their respective means, there is no intercept parameter in the **B** vectors. The $\hat{\mathbf{Y}}$ vectors are centred, as is always the case in ordinary linear regression. If $m = n$, **X** is square; in that case, the multiple regressions always explain the variables in matrix **Y** entirely, so that $\hat{\mathbf{Y}} = \mathbf{Y}$. Using property 5 of matrix inverses (Section 2.8), one can indeed verify that eq. 11.6 gives $\hat{\mathbf{Y}} = \mathbf{Y}$ when **X** is square.

2) The covariance matrix corresponding to the table of fitted values $\hat{\mathbf{Y}}$ is computed from eq. 4.6:

$$\mathbf{S}_{\hat{\mathbf{Y}}'\hat{\mathbf{Y}}} = [1/(n-1)]\,\hat{\mathbf{Y}}'\hat{\mathbf{Y}} \qquad \textbf{(11.7)}$$

Replacing $\hat{\mathbf{Y}}$ by the expression from eq. 11.6, eq. 11.7 becomes:

$$\mathbf{S}_{\hat{\mathbf{Y}}'\hat{\mathbf{Y}}} = [1/(n-1)]\,\mathbf{Y}'\mathbf{X}[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{X}\,[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Y} \qquad \textbf{(11.8)}$$

This equation reduces to:

$$\mathbf{S}_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}} = \mathbf{S}_{\mathbf{YX}} \mathbf{S}^{-1}_{\mathbf{XX}} \mathbf{S'}_{\mathbf{YX}} \qquad \textbf{(11.9)}$$

where $\mathbf{S_{YY}}$ is the $(p \times p)$ covariance matrix among the response variables, $\mathbf{S_{XX}}$ the $(m \times m)$ covariance matrix among the regressors (it is actually a matrix $\mathbf{R_{XX}}$ if all the $\mathbf{X}$ variables have been standardized, as suggested above), and $\mathbf{S_{YX}}$ is the $(p \times m)$ covariance matrix among the variables of the two sets; its transpose $\mathbf{S'_{YX}} = \mathbf{S_{XY}}$ is of size $(m \times p)$. If the $\mathbf{Y}$ variables had also been standardized, this equation would read $\mathbf{R_{YX}}\,\mathbf{R}^{-1}_{\mathbf{XX}}\,\mathbf{R'_{YX}}$, which is the equation for the coefficient of multiple determination (eq. 4.31).

3) The table of fitted values $\hat{\mathbf{Y}}$ is subjected to principal component analysis to reduce the dimensionality of the solution. This corresponds to solving the eigenvalue problem:

$$(\mathbf{S}_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}} - \lambda_k \mathbf{I})\,\mathbf{u}_k = \mathbf{0} \qquad \textbf{(11.10)}$$

which, using eq. 11.9, translates into:

$$(\mathbf{S_{YX}}\,\mathbf{S}^{-1}_{\mathbf{XX}}\,\mathbf{S'_{YX}} - \lambda_k \mathbf{I})\,\mathbf{u}_k = \mathbf{0} \qquad \textbf{(11.11)}$$

This is the equation for redundancy analysis (eq. 11.3); it may also be obtained from the equation for canonical correlation analysis (eq. 11.22), by defining $\mathbf{S}_{11} = \mathbf{S_{YY}} = \mathbf{I}$ (Rao, 1973; ter Braak, 1987c). Different programs may express the eigenvalues in different ways: raw eigenvalues, fraction of total variance in matrix $\mathbf{Y}$, or percentage; see Tables 11.2 and 11.4 for examples.

The matrix containing the normalized canonical eigenvectors $\mathbf{u}_k$ is called $\mathbf{U}$. The eigenvectors give the contributions of the descriptors of $\hat{\mathbf{Y}}$ to the various canonical axes. Matrix $\mathbf{U}$, of size $(p \times p)$, contains only $\min[p, m, n-1]$ eigenvectors with non-zero eigenvalues, since the number of canonical eigenvectors cannot exceed the minimum of p, m and $(n-1)$:

- It cannot exceed p which is the size of the reference space of matrix $\mathbf{Y}$. This is obvious in multiple regression, where matrix $\mathbf{Y}$ contains a single variable; the ordination given by the fitted values $\hat{y}$ is, consequently, one-dimensional.

- It cannot exceed m which is the number of variables in $\mathbf{X}$. Consider an extreme example: if $\mathbf{X}$ contains a single explanatory variable ($m = 1$), regressing all p variables in $\mathbf{Y}$ on this single regressor produces p fitted vectors $\hat{y}$ which all point in the same direction of the space; a principal component analysis of matrix $\hat{\mathbf{Y}}$ of these fitted vectors can only produce one common (canonical) variable.

- It cannot exceed $(n-1)$ which is the maximum number of dimensions required to represent n points in Euclidean space.

The canonical coefficients in the normalized matrix **U** give the contributions of the variables of $\hat{\mathbf{Y}}$ to the canonical axes. They should be interpreted as in PCA. For biplots (discussed below), matrix **U** can be rescaled in such a way that the length of each eigenvector is $\sqrt{\lambda_k}$, using eq. 9.9.

4) The ordination of objects in the space of the response variables **Y** can be obtained directly from the centred matrix **Y**, using the standard equation for principal components (eq. 9.4) and matrix **U** of the eigenvectors $\mathbf{u}_k$ found in eq. 11.11:

$$\mathbf{F} = \mathbf{Y}\mathbf{U} \qquad \textbf{(11.12)}$$

Site scores

The ordination vectors (columns of **F**) defined in eq. 11.12 are called the vectors of "site scores". They have variances that are close, but not equal to the corresponding eigenvalues. How to represent matrix **F** in biplot diagrams is discussed in point 8 (below).

5) Likewise, the ordination of objects in space **X** is obtained as follows:

$$\mathbf{Z} = \hat{\mathbf{Y}}\mathbf{U} = \mathbf{X}\mathbf{B}\mathbf{U} \qquad \textbf{(11.13)}$$

Fitted site scores

As stated above, the vectors in matrix $\hat{\mathbf{Y}}$ are centred on their respective means. The right-hand part of eq. 11.13, obtained by replacing $\hat{\mathbf{Y}}$ by it value in eq. 11.5, shows that this ordination is a linear combinations of the **X** variables. For that reason, these ordination vectors (columns of matrix **Z**) are also called "fitted site scores", or "sample scores which are linear combinations of environmental variables" in program CANOCO. The ordination vectors, as defined in eq. 11.13, have variances equal to the corresponding eigenvalues. The representation of matrix **Z** in biplot diagrams is discussed in point 8 (below).

The "site scores" of eq. 11.12 are obtained by projecting the original data (matrix **Y**) onto axis *k*; they approximate the observed data, which contain residuals ($\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{Y}_{res}$, Fig. 11.2). On the other hand, the "fitted site scores" of eq. 11.13 are obtained by projecting the fitted values of the multiple regressions (matrix $\hat{\mathbf{Y}}$) onto axis *k*; they approximate the fitted data. Either set may be used in biplots. The practical difference between "site scores" and "fitted site scores" is further discussed in the second example below and in the numerical example of Section 13.4.

6) The correlation r_k between the ordination vectors in spaces **Y** (from eq. 11.12) and **X** (from eq. 11.13) for dimension *k* is called the "species-environment correlation" in program CANOCO. It measures how strong the relationship is between the two data sets, as expressed by each canonical axis *k*. It should be interpreted with caution because a canonical axis with high species-environment correlation may explain but a small fraction of the variation in **Y**, which is given by the amount (or proportion) of variance of matrix **Y** explained by each canonical axis; see example in Table 11.2.

7) The last important information needed for interpretation is the contribution of the explanatory variables **X** to the canonical ordination axes. Either the regression or the correlation coefficients may be considered:

- Matrix **C** of the canonical coefficients,

$$\mathbf{C} = \mathbf{B}\,\mathbf{U} \qquad \textbf{(11.14)}$$

gives directly the weights of the explanatory variables **X** in the formation of the matrix of fitted site scores. The ordination of objects in the space of the explanatory variables can be found directly by computing **XC**; these vectors of site scores are the same as in eq. 11.13. The coefficients in the columns of matrix **C** are identical to the regression coefficients of the ordination scores from eq. 11.13 on the matrix of standardized explanatory variables **X**; they may thus be interpreted in the same way.

- Correlations may also be computed between the variables in **X**, on the one hand, and the ordination vectors, in either space **Y** (from eq. 11.12) or space **X** (from eq. 11.13), on the other. The correlations between **X** and the ordination vectors in space **X** are used to represent the explanatory variables in biplots.

Biplot

8) In RDA, *biplot diagrams* may be drawn that contain two sets of points, as in PCA (Subsection 9.1.4), or three sets: site scores (matrices **F** or **Z**, from eqs. 11.12 and 11.13), response variables from **Y,** and explanatory variables from **X**. Each pair of sets of points forms a biplot. Biplots help interpret the ordination of objects in terms of **Y** and **X**. When there are too many objects, or too many variables in **Y** or **X**, separate ordination diagrams may be drawn and presented side by side. The construction of RDA biplot diagrams is explained in detail in ter Braak (1994); his conclusions are summarized here. As in PCA, two main types of scalings may be used (Table 9.2):

Scalings in RDA

- RDA scaling type 1 — The eigenvectors in matrix **U**, representing the scores of the response variables along the canonical axes, are scaled to lengths 1[*]. The site scores in space **X** are obtained from equation $\mathbf{Z} = \hat{\mathbf{Y}}\mathbf{U}$ (eq. 11.13); these vectors have variances equal to λ_k. The site scores in space **Y** are obtained from equation $\mathbf{F} = \mathbf{YU}$; the variances of these vectors are usually slightly larger than λ_k because **Y** contains both the fitted and residual components and has thus more total variance than $\hat{\mathbf{Y}}$. Matrices **Z** and **U**, or **F** and **U**, can be used together in biplots because the products of the eigenvectors with the site score matrices reconstruct the original matrices perfectly:

[*] In CANOCO 3.1, RDA scaling –1 produces a matrix **U** with vectors ("species scores") scaled to lengths $\sqrt{n}$ (or $\sqrt{p}$ in CANOCO 4.0), instead of 1, if all species and site weights are equal. In both versions of CANOCO, the site scores in space **X** (matrix **Z**) are scaled to lengths $\sqrt{n\lambda_k}$ (or, in other words, to sums of squares of $n\lambda_k$); the site scores in space **Y** (matrix **F**) have lengths slightly larger than $\sqrt{n\lambda_k}$. For RDA, CANOCO expresses the eigenvalues as fractions of the total variance in **Y**. As a result, site scores in matrices **F** and **Z**, as described in the present Section (z_{here}), are related to site scores given by CANOCO (z_{CANOCO}) through the formula: $z_{\text{CANOCO}} = z_{\text{here}} \sqrt{n / ((n-1)\,\text{Total variance in } \mathbf{Y})}$. Changing the scaling of species and site score vectors by any multiplicative constant does not change the interpretation of a biplot.

$\mathbf{ZU}' = \hat{\mathbf{Y}}$ and $\mathbf{FU}' = \mathbf{Y}$, as in PCA (Subsection 9.1.4). A quantitative explanatory variable **x** may be represented in the biplot using the correlations of **x** with the fitted site scores. Each correlation is multiplied by $\sqrt{\lambda_k / \text{Total variance in } \mathbf{Y}}$ where λ_k is the eigenvalue of the corresponding axis *k*; this correction accounts for the fact that, in this scaling, the variances of the site scores differ among axes. The correlations were obtained in calculation step 7 above.

The consequences of this scaling, for PCA, are summarized in the right-hand column of Table 9.2. This scaling, called *distance biplot*, allows the interpretation to focus on the ordination of objects because the distances among objects approximate their Euclidean distances in the space of response variables (matrix **Y**).

Distance biplot

The main features of a distance biplot are the following: (1) Distances among objects in a biplot are approximations of their Euclidean distances. (2) Projecting an object at right angle on a response variable **y** approximates the value of the object along that variable, as in Fig. 9.3a. (3) The angles among variables **y** are meaningless. (4) The angles between variables **x** and **y** in the biplot reflect their correlations. (5) Binary explanatory **x** variables may be represented as the centroids of the objects possessing state "1" for that variable. Examples are given in Subsection 2. Since a centroid represents a "mean object", its relationship to a variable **y** is found by projecting it at right angle on the variable, as for an object. Distances among centroids, and between centroids and individual objects, approximate Euclidean distances.

• RDA scaling type 2 — Alternatively, one obtains response variable scores by rescaling the eigenvectors in matrix **U** to lengths $\sqrt{\lambda_k}$, using the transformation $\mathbf{U\grave{O}}^{1/2*}$. The site scores in space **X** obtained for scaling 1 (eq. 11.13) are rescaled to unit variances using the transformation $\mathbf{Z\grave{O}}^{-1/2}$. The site scores in space **Y** obtained for scaling 1 are rescaled using the transformation $\mathbf{F\,\grave{O}}^{-1/2}$; the variances of these vectors are usually slightly larger than 1 for the reason explained in the case of scaling 1. Matrices **Z** and **U**, or **F** and **U**, as rescaled here, can be used together in biplots because the products of the eigenvectors with the site score matrices reconstruct the original matrices perfectly: $\mathbf{ZU}' = \hat{\mathbf{Y}}$ and $\mathbf{FU}' = \mathbf{Y}$, as in PCA (Subsection 9.1.4). A quantitative explanatory variable **x** may be represented in the biplot using the correlations of **x** with the fitted site scores, obtained in calculation step 7 above.

---

* In CANOCO 3.1, RDA scaling –2 produces a matrix **U** with vectors ("species scores") scaled to lengths $\sqrt{n\lambda_k}$ (or $\sqrt{p\lambda_k}$ in CANOCO 4.0), instead of $\sqrt{\lambda_k}$, if all species and site weights are equal. For RDA, CANOCO expresses the eigenvalues as fractions of the total variance in **Y**. As a result, the values in matrix **U** as described here (u_{here}) are related to the "species scores" of CANOCO 3.1 ($u_{\text{CANOCO 3.1}}$) through the formula: $u_{\text{CANOCO 3.1}} = u_{\text{here}} \sqrt{n / \text{Total variance in } \mathbf{Y}}$, or $u_{\text{CANOCO 4.0}} = u_{\text{here}} \sqrt{p / \text{Total variance in } \mathbf{Y}}$ in CANOCO 4.0. In both versions of CANOCO, the site scores in space **X** (matrix **Z**) are scaled to lengths $\sqrt{n}$ instead of $\sqrt{n-1}$; the site scores in space **Y** (matrix **F**) have lengths slightly larger than $\sqrt{n}$. Site scores in matrices **F** and **Z**, as described in the present Section (z_{here}), are related to site scores given by CANOCO (z_{CANOCO}) through the formula: $z_{\text{CANOCO}} = z_{\text{here}} \sqrt{n/(n-1)}$. Changing the scaling of species and site score vectors by any multiplicative constant does not change the interpretation of a biplot.

The consequences of this scaling, for PCA, are summarized in the central column of Table 9.2. This scaling, called *correlation biplot*, is appropriate to focus on the relationships among response variables (matrix **Y**).

Correlation biplot

The main features of a correlation biplot are the following: (1) Distances among objects in the biplot *are not* approximations of their Euclidean distances. (2) Projecting an object at right angle on a response variable **y** approximates the value of the object along that variable. (3) The angles between variables (from sets **X** and **Y**) in the biplot reflect their correlations. (4) Projecting an object at right angle on a variable **x** approximates the value of the object along that variable. (5) Binary explanatory variables may be represented as described above. Their interpretation is done in the same way as in scaling type 1, except for the fact that the distances in the biplot among centroids, and between centroids and individual objects, do not approximate Euclidean distances.

The type of scaling depends on the emphasis one wants to give to the biplot, i.e. display of distances among objects or of correlations among variables. When most explanatory variables are binary, scaling type 1 is probably the most interesting; when most of the variables in set **X** are quantitative, one may prefer scaling type 2. When the first two eigenvalues are nearly equal, both scalings lead to nearly the same biplot.

9) Redundancy analysis usually does not completely explain the variation in the response variables (matrix **Y**). During the regression step (Fig. 11.2), regression residuals may be computed for each variable **y**; the residuals are obtained as the difference between observed values y_{ij} and the corresponding fitted values $\hat{y}_{ij}$ in matrix $\hat{\mathbf{Y}}$. The matrix of residuals ($\mathbf{Y}_{\text{res}}$ in Fig. 11.2) is also a matrix of size $(n \times p)$. Residuals may be analysed by principal component analysis, leading to $\min[p, n-1]$ non-canonical eigenvalues and eigenvectors (Fig. 11.2, bottom). So, the full analysis of matrix **Y** (i.e. the analysis of fitted values and residuals) may lead to more eigenvectors than a principal component analysis of matrix **Y**: there is a maximum of $\min[p, m, n-1]$ non-zero canonical eigenvalues and corresponding eigenvectors, plus a maximum of $\min[p, n-1]$ non-canonical eigenvalues and eigenvectors, the latter being computed from the matrix of residuals (Table 11.1). When the variables in **X** are good predictors of the variables in **Y**, the canonical eigenvalues may be larger than the first non-canonical eigenvalues, but this need not always be the case. If the variables in **X** are not good predictors of **Y**, the first non-canonical eigenvalues, computed on the residuals, may be larger than their canonical counterparts.

In the trivial case where **Y** contains a single response variable, redundancy analysis is nothing but multiple linear regression analysis.

2 — Numerical examples

As a first example, consider again the data set presented in Table 10.5. The first five variables are assembled into matrix **Y** and the three spatial variables make up matrix **X**. Calculations performed as described above, or using the iterative algorithm

Table 11.1 Maximum number of non-zero eigenvalues and corresponding eigenvectors that may be obtained from canonical analysis of a matrix of response variables $\mathbf{Y}(n \times p)$ and a matrix of explanatory variables $\mathbf{X}(n \times m)$ using redundancy analysis (RDA) or canonical correspondence analysis (CCA).

	Canonical eigenvalues and eigenvectors	Non-canonical eigenvalues and eigenvectors
RDA	$\min[p, m, n-1]$	$\min[p, n-1]$
CCA	$\min[(p-1), m, n-1]$	$\min[(p-1), n-1]$

described in the next subsection, lead to the same results (Table 11.2). There are min[5, 3, 19] = 3 canonical eigenvectors in this example and 5 non-canonical PCA axes computed from the residuals. This is a case where the first non-canonical eigenvalue is larger than any of the canonical eigenvalues. The ordination of objects along the canonical axes (calculation steps 4 and 5 in the previous Subsection) as well as the contribution of the explanatory variables to the canonical ordination axes (calculation step 6) are not reported, but the correlations between these two sets of ordinations are given in the Table; they are rather weak. The sum of the three canonical eigenvalues accounts for only 32% of the variation in response matrix **Y**.

A second example has been constructed to illustrate the calculation and interpretation of redundancy analysis. Imagine that fish have been observed at 10 sites along a transect running from the beach of a tropical island, with water depths going from 1 to 10 m (Table 11.3). The first three sites are on sand and the others alternate between coral and "other substrate". The first six species avoid the sandy area, possibly because little food is available there, whereas the last three are ubiquitous. The sums of abundances for the 9 species are in the last row of the Table. Species 1 to 6 come in three successive pairs, with distributions forming opposite gradients of abundance between sites 4 and 10. Species 1 and 2 are not associated to a single type of substrate. Species 3 and 4 are found in the coral areas only while species 5 and 6 are found on other substrates only (coral debris, turf, calcareous algae, etc.). The distributions of abundance of the ubiquitous species (7 to 9) have been produced using a random number generator, fitting the frequencies to a predetermined sum; these species will only be used to illustrate CCA in Section 11.2.

RDA was computed using the first six species as matrix **Y**, despite the fact that CCA (Subsection 11.2) is probably more appropriate for these data. Comparison of Tables 11.4 and 11.5, and of Figs. 11.3 and 11.5, allows, to a certain extent, a comparison of the two methods. The analysis was conducted on centred **y** variables because species abundances do not require standardization. When they are not

Table 11.3 Artificial data set representing observations (e.g. fish abundances) at 10 sites along a tropical reef transect. The variables are further described in the text.

Site No.	Sp. 1	Sp. 2	Sp. 3	Sp. 4	Sp. 5	Sp. 6	Sp. 7	Sp. 8	Sp. 9	Depth (m)	Substrate type Coral	Sand	Other
1	1	0	0	0	0	0	2	4	4	1	0	1	0
2	0	0	0	0	0	0	5	6	1	2	0	1	0
3	0	1	0	0	0	0	0	2	3	3	0	1	0
4	11	4	0	0	8	1	6	2	0	4	0	0	1
5	11	5	17	7	0	0	6	6	2	5	1	0	0
6	9	6	0	0	6	2	10	1	4	6	0	0	1
7	9	7	13	10	0	0	4	5	4	7	1	0	0
8	7	8	0	0	4	3	6	6	4	8	0	0	1
9	7	9	10	13	0	0	6	2	0	9	1	0	0
10	5	10	0	0	2	4	0	1	3	10	0	0	1
Sum	60	50	40	30	20	10	45	35	25				

Results of the analysis are presented in Table 11.4; programs such as CANOCO provide more output tables than presented here. The data could have produced 3 canonical axes and up to 6 non-canonical eigenvectors. In this example, only 4 of the 6 non-canonical axes had variance larger than 0. An overall test of significance (Subsection 11.3.2) showed that the canonical relationship between matrices **X** and **Y** is very highly significant ($p = 0.001$ after 999 permutations; permutation of residuals using CANOCO). The canonical axes explain 66%, 22% and 8% of the response table's variance, respectively; they are all significant ($p < 0.05$) and display strong species-environment correlations ($r = 0.999$, 0.997, and 0.980, respectively).

In Table 11.4, the eigenvalues are first given with respect to the total variance in matrix **Y**, as is customary in principal component analysis. They are also presented as proportions of the total variance in **Y** as is the practice in program CANOCO in the case of PCA and RDA. The species and sites are scaled for a distance biplot (RDA scaling type 1, Subsection 11.1.1). The eigenvectors (called "species scores" in CANOCO) are normalized to length 1. The site scores (matrix **F**) are obtained from eq. 11.12. They provide the ordination of the objects in the space of the original matrix **Y**. These ordination axes are not orthogonal to one another because matrix **Y** contains the "residual" components of the multiple regressions (Fig. 11.2). The "site scores that are linear combinations of the environmental variables", or "fitted site sores" (matrix **Z**, not printed in Table 11.4), are obtained from eq. 11.13. They provide the ordination of the objects in the space of matrix $\hat{\mathbf{Y}}$ which contains the fitted values of the multiple regressions (Fig. 11.2). These ordination axes are orthogonal to one another because

Table 11.4 Results of redundancy analysis of data in Table 11.3 (selected output). Matrix **Y**: species 1 to 6. Matrix **X**: depth and substrate classes.

	Canonical axes			Non-canonical axes			
	I	II	III	IV	V	VI	VII
Eigenvalues (with respect to total variance in **Y** = 112.88889)							
	74.52267	24.94196	8.87611	4.18878	0.31386	0.03704	0.00846
Fraction of total variance in **Y** (these are the eigenvalues of program CANOCO for RDA)							
	0.66014	0.22094	0.07863	0.03711	0.00278	0.00033	0.00007
Cumulative fraction of total variance in **Y** accounted for by axes 1 to *k*							
	0.66014	0.88108	0.95971	0.99682	0.99960	0.99993	1.00000
Normalized eigenvectors ("species scores"): mat. **U** for the canonical and $\mathbf{U_{res}}$ for the non-canonical portions							
Species 1	0.30127	–0.64624	–0.39939	–0.00656	–0.40482	0.70711	–0.16691
Species 2	0.20038	–0.47265	0.74458	0.00656	0.40482	0.70711	0.16690
Species 3	0.74098	0.16813	–0.25690	–0.68903	–0.26668	0.00000	0.67389
Species 4	0.55013	0.16841	0.26114	0.58798	0.21510	0.00000	0.68631
Species 5	–0.11588	–0.50594	–0.29319	0.37888	–0.66624	0.00000	0.12373
Species 6	–0.06292	–0.21535	0.25679	–0.18944	0.33312	0.00000	–0.06187
Site scores ("sample scores"): matrices **F** for the canonical and non-canonical portions, eqs. 11.12 and 9.4							
Site 1	–6.82791	5.64392	–1.15219	0.24712	1.14353	0.23570	0.01271
Site 2	–7.12919	6.29016	–0.75280	0.00000	0.00000	–0.47140	0.00000
Site 3	–6.92880	5.81751	–0.00823	–0.24712	–1.14353	0.23570	–0.01271
Site 4	–4.00359	–6.97190	–4.25652	2.14250	–0.28230	0.00000	0.00141
Site 5	13.63430	0.85534	–3.96242	–3.80923	–0.14571	0.00000	0.10360
Site 6	–4.03654	–5.82821	–1.12541	0.71417	–0.09410	0.00000	0.00047
Site 7	12.11899	1.03525	0.13651	0.22968	0.08889	0.00000	–0.22463
Site 8	–4.06949	–4.68452	2.00570	–0.71417	0.09410	0.00000	–0.00047
Site 9	11.34467	1.38328	3.97855	3.57956	0.05682	0.00000	0.12103
Site 10	–4.10243	–3.54082	5.13681	–2.14250	0.28230	0.00000	–0.00141
Correlations of environmental variables with site scores from eq. 11.12							
Depth	0.42204	–0.55721	0.69874				
Coral	0.98708	0.15027	0.01155				
Sand	–0.55572	0.81477	–0.14471				
Other subs.	–0.40350	–0.90271	0.12456				
Biplot scores of environmental variables							
Depth	0.34340	–0.26282	0.20000				
Coral	0.80314	0.07088	0.00330				
Sand	–0.45216	0.38431	–0.04142				
Other subs.	–0.32831	–0.42579	0.03565				
Centroids of sites with code "1" for BINARY environmental variables, in ordination diagram							
Coral	12.36599	1.09129	0.05088				
Sand	–6.96197	5.91719	–0.63774				
Other subs.	–4.05301	–5.25636	0.44014				

the eigenanalysis (PCA in Fig. 11.2) has been conducted on matrix $\hat{\mathbf{Y}}$. Both the "site scores" (matrix **F**) and "fitted site scores" (matrix **Z**) may be used in RDA biplots.*

Correlations of the environmental variables with the ordination vectors can be obtained in two forms: either with respect to the "site scores" (eq. 11.12) or with respect to the "fitted site scores" (eq. 11.13). The latter set of correlations is used to draw biplots containing the sites as well as the variables from **Y** and **X** (Fig. 11.3). There were three binary variables in Table 11.3. Each such variable may be represented by the centroid of the sites possessing state "1" for that variable (or else, the centroid of the sites possessing state "0"). These three variables are represented by both arrows (correlations) and symbols (centroids) in Fig. 11.3 to show the difference between these representations; in real-case studies, one chooses one of the representations.

The following question may arise when the effect of some environmental variables on the dependent variables **Y** is already well known (e.g. the effect of altitude on vegetation along a mountain slope, or the effect of depth on plankton assemblages): what would the *residual* ordination of sites (or the *residual* correlations among variables) be like if one could control for the linear effect of such well-known environmental variables? An approximate answer may be obtained by looking at the structure of the *residuals* obtained by regressing the original variables on the variables representing the well-known factors. With the present data set, for instance, one could examine the residual structure, after controlling for depth and substrate, by plotting ordination biplots of the *non-canonical axes* in Table 11.4. These axes correspond to a PCA of the table of residual values of the multiple regressions (Fig. 11.2).

3 — Algorithms

There are different ways of computing RDA. One may go through the multiple regression and principal component analysis steps described in Fig. 11.2, or calculate the matrix corresponding to $\mathbf{S_{YX}}\,\mathbf{S^{-1}_{XX}}\,\mathbf{S'_{YX}}$ in eq. 11.3 and decompose it using a standard eigenvalue-eigenvector algorithm.

Alternatively, ter Braak (1987c) suggested to modify his iterative algorithm for principal component analysis (Table 9.5), by incorporating a regression analysis at the end of each iteration, as illustrated in Fig. 11.4. Because it would be cumbersome to repeat a full multiple regression calculation at the end of each iteration and for each

* To obtain a distance biplot based upon the covariance matrix using program CANOCO (version 3 or later), one should centre the response variables (no standardization) and emphasize the ordination of sites by choosing scaling –1 in the "long dialogue" option. In the Windows version of CANOCO 4.0, focus on inter-site distances and do not post-transform the species scores. CANOCO prints the eigenvalues as proportions of the total variation in matrix **Y**. The scalings of eigenvalues and eigenvectors produced by CANOCO are described in the footnotes of Subsection 11.1.1. Changing the scaling of species and site score vectors by any multiplicative constant does not change the interpretation of a biplot.

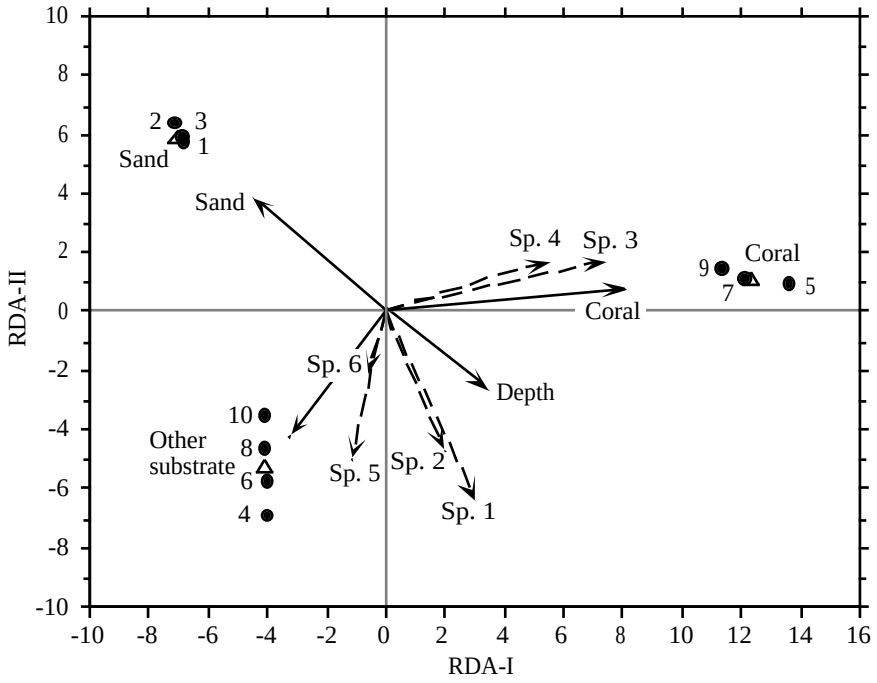


Figure 11.3 RDA ordination biplot of the artificial data presented in Table 11.3; the numerical results of the analysis are in Table 11.4. Dots are the sampling sites; numbers represent both the site number identifiers and depths (in m). Dashed arrows are the species. Full-line arrows represent the "biplot scores of environmental variables". The lengths of all arrows have been multiplied by 10 for clarity of the diagram. The "centroids of sites with code 1 for [the three] binary environmental variables" are represented by triangles. Binary environmental variables are usually represented by *either* arrows *or* symbols, not both as in this diagram.

canonical eigenvector, a short cut can be used. Vector **b** of regression coefficients is obtained from eq. 2.19:

$$\mathbf{b} = [\mathbf{X}'\mathbf{X}]^{-1}\,[\mathbf{X}'\mathbf{y}]$$

Only the $[\mathbf{X}'\mathbf{y}]$ portion must be recomputed during each iteration of the estimation of the canonical eigenvectors; the $[\mathbf{X}'\mathbf{X}]^{-1}$ part, which is the most cumbersome to calculate, is constant during the whole redundancy analysis run so that it needs to be computed only once.

The iterative procedure presents two advantages: (1) with large problems, one is satisfied, in most instances, with computing a few axes only, instead of having to estimate all eigenvalues and eigenvectors. The iterative procedure was developed to do

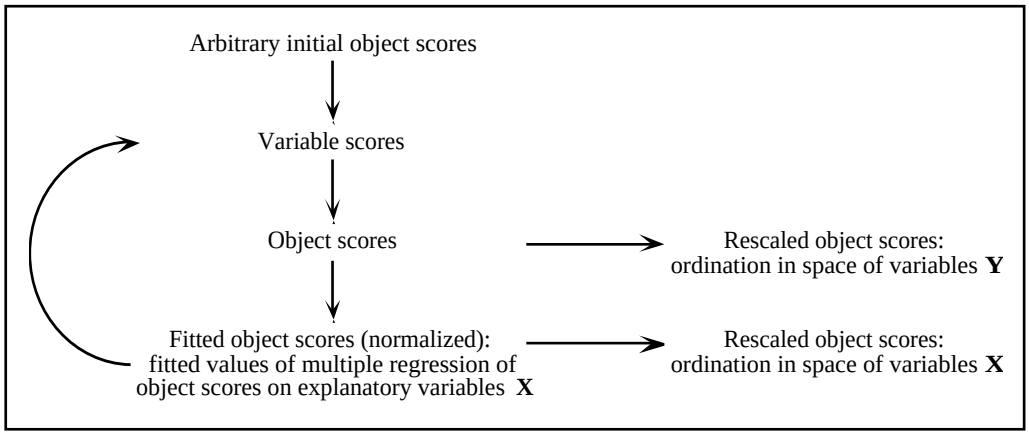


Figure 11.4 Two-way weighted summation algorithm (from Table 9.5), modified to compute redundancy analysis. Two types of ordinations are produced, one in the space of the **Y** variables and the other in the space of the **X** variables. Translated from Borcard & Buttler (1997).

that. (2) With smaller problems, the canonical *and* non-canonical axes can all be computed at once; one does not have to carry out a separate calculation on the matrix of residuals to obtain the non-canonical axes. The main disadvantage of the procedure is the possibility of numerical instability *when a large number of successive axes are computed*. For large problems in which all the canonical axes are needed, this procedure also requires more calculation time than regular eigenanalysis.

11.2 Canonical correspondence analysis (CCA)

Canonical correspondence analysis is a canonical ordination method developed by ter Braak (1986, 1987a, 1987c) and implemented in the program CANOCO (ter Braak, 1988b, 1988c, 1990; ter Braak & Smilauer, 1998). It is the canonical form of correspondence analysis. Any data table that could be subjected to correspondence analysis forms a suitable *response matrix* **Y** for CCA; this is the case, in particular, for species presence-absence or abundance tables (Section 9.4).

1 — The algebra of canonical correspondence analysis

The mathematics of CCA is essentially the same as that of redundancy analysis. Differences involve the diagonal matrices of row totals $\mathbf{D}(f_{i+})$ and row relative frequencies $\mathbf{D}(p_{i+})$, as defined in Section 9.4 for correspondence analysis; f_{i+} is the sum of values in row i of matrix **Y** whereas p_{i+} is f_{i+} divided by the grand total f_{++} of all values in **Y**.

Inflated data matrix

The calculations are modified in such a way as to simulate an analysis carried out on *inflated data matrices* $\mathbf{Y}_{infl}$ and $\mathbf{X}_{infl}$ constructed in a way similar to the inflated data table of Subsection 9.4.4. Assume that **Y** contains species presence-absence or abundance data and **X** contains environmental data. A *single species presence*, from the original table of species abundances, is placed in each row of $\mathbf{Y}_{infl}$. An inflated matrix $\mathbf{Y}_{infl}$ usually has many more rows than matrix **Y**. In the corresponding inflated matrix $\mathbf{X}_{infl}$, the row vectors of environmental data are repeated as required to make every species presence (in $\mathbf{Y}_{infl}$) face a copy of the appropriate vector of environmental data (in $\mathbf{X}_{infl}$). Modifications to the RDA algorithm are the following:

- The dependent data table is not matrix **Y** centred by variables (columns) as in RDA. CCA uses matrix $\overline{\mathbf{Q}}$ of the contributions to chi-square, also used in correspondence analysis. $\overline{\mathbf{Q}}$ is derived from matrix **Y** using eq. 9.32. Matrix **Q** of the relative frequencies is also computed ($\mathbf{Q} = (1/f_{++})\mathbf{Y}$); it is used in the scaling operations.

- Matrix **X** is standardized using weights $\mathbf{D}(f_{i+})$. To achieve this, compute the mean and standard deviation for each column of the inflated matrix $\mathbf{X}_{infl}$, which contains f_{++} rows, and use them to standardize the environmental data. Use the maximum likelihood estimator for the variance instead of the usual estimator (eq. 4.3); in other words, divide the sum of squared deviations from the mean by the number of rows of matrix $\mathbf{X}_{infl}$ (which is equal to f_{++}), instead of the number of rows minus 1.

- Weighted multiple regression is used instead of a conventional multiple regression. The weights, given by diagonal matrix $\mathbf{D}(p_{i+})^{1/2}$, are applied to matrix **X** everywhere it occurs in the multiple regression equations, which become:

$$\mathbf{B} = [\mathbf{X}' \mathbf{D}(p_{i+}) \mathbf{X}]^{-1} \mathbf{X}' \mathbf{D}(p_{i+})^{1/2} \overline{\mathbf{Q}}$$

and

$$\hat{\mathbf{Y}} = \mathbf{D}(p_{i+})^{1/2} \mathbf{X} \mathbf{B}$$

The equation for computing $\hat{\mathbf{Y}}$ is then:

$$\hat{\mathbf{Y}} = \mathbf{D}(p_{i+})^{1/2} \mathbf{X} [\mathbf{X}' \mathbf{D}(p_{i+}) \mathbf{X}]^{-1} \mathbf{X}' \mathbf{D}(p_{i+})^{1/2} \overline{\mathbf{Q}} \qquad \textbf{(11.15)}$$

The matrix of residuals is computed as $\mathbf{Y_{res}} = \overline{\mathbf{Q}} - \hat{\mathbf{Y}}$. This is the equivalent, for CCA, of the equation $\mathbf{Y_{res}} = \mathbf{Y} - \hat{\mathbf{Y}}$ used in Fig. 11.2 for RDA.

- Eigenvalue decomposition (eqs. 11.10 and 11.11) is carried out on matrix $\mathbf{S}_{\hat{\mathbf{Y}}'\hat{\mathbf{Y}}}$ which, in this case, is simply the matrix of sums of squares and cross products, without division by the number of degrees of freedom — as in correspondence analysis:

$$\mathbf{S}_{\hat{\mathbf{Y}}'\hat{\mathbf{Y}}} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} \qquad \textbf{(11.16)}$$

One can show that $\mathbf{S}_{\hat{\mathbf{Y}}\,\hat{\mathbf{Y}}}$, computed as described, is equal to $\mathbf{S}_{\bar{\mathbf{Q}}\mathbf{X}}\mathbf{S}_{\mathbf{XX}}^{-1}\mathbf{S}'_{\bar{\mathbf{Q}}\mathbf{X}}$ (eq. 11.3 and 11.11) if the covariance matrices $\mathbf{S}_{\bar{\mathbf{Q}}\mathbf{X}}$ and $\mathbf{S}_{\mathbf{XX}}$ are computed with weights on $\mathbf{X}$, given by matrix $\mathbf{D}(p_{i+})^{1/2}$, and without division by the number of degrees of freedom.

With these modifications, CCA is computed using eq. 11.3, obtaining matrices $\mathbf{\Lambda}$ of eigenvalues and $\mathbf{U}$ of eigenvectors. Canonical correspondence analysis is thus a weighted form of redundancy analysis, applied to dependent matrix $\bar{\mathbf{Q}}$. It approximates chi-square distances among the rows (objects) of the dependent data matrix, subject to the constraint that the canonical ordination vectors be maximally related to weighted linear combinations of the explanatory variables. The equations are also described in Section 5.9.5 of ter Braak (1987c). The method is perfectly suited to analyse the relationships between species presence/absence or abundance data matrices and tables of environmental variables. The number of canonical and non-canonical axes expected from the analysis are given in Table 11.1. Tests of significance are available, in CCA and RDA, for the total canonical variation and for individual eigenvalues (Subsection 11.3.2).

- The normalized matrix $\hat{\mathbf{U}}$ is obtained using eq. 9.38:

$$\hat{\mathbf{U}} = \bar{\mathbf{Q}}\,\mathbf{U}\mathbf{\Lambda}^{-1/2}$$

In CCA, matrix $\hat{\mathbf{U}}$ as defined here does not contain the loadings of the rows of $\hat{\mathbf{Y}}$ on the canonical axes. It contains instead the loadings of the rows of $\bar{\mathbf{Q}}$ on the ordination axes, as in CA. It will be used to find the site scores (matrices $\mathbf{F}$ and $\hat{\mathbf{V}}$) in the space of the original variables $\mathbf{Y}$. The site scores in the space of the fitted values $\hat{\mathbf{Y}}$ will be found using $\mathbf{U}$ instead of $\hat{\mathbf{U}}$.

Scalings in CCA

- Matrix $\mathbf{V}$ of species scores (for scaling type 1) and matrix $\hat{\mathbf{V}}$ of site scores (for scaling type 2) are obtained from $\mathbf{U}$ and $\hat{\mathbf{U}}$ using the transformations described for correspondence analysis (Subsection 9.4.1):

eq. 9.41 (species scores, scaling 1): $\mathbf{V} = \mathbf{D}(p_{+j})^{-1/2}\mathbf{U}$

and eq. 9.42 (site scores, scaling 2): $\hat{\mathbf{V}} = \mathbf{D}(p_{i+})^{-1/2}\,\hat{\mathbf{U}}$

or combining eqs. 9.38 and 9.42: $\hat{\mathbf{V}} = \mathbf{D}(p_{i+})^{-1/2}\,\bar{\mathbf{Q}}\,\mathbf{U}\mathbf{\Lambda}^{-1/2}$

Scalings 1 and 2 are the same as in correspondence analysis (Subsection 9.4.1). Matrices $\mathbf{F}$ (site scores for scaling type 1) and $\hat{\mathbf{F}}$ (species scores for scaling type 2) are found using eqs. 9.43a and 9.44a:

$$\mathbf{F} = \hat{\mathbf{V}}\,\mathbf{\Lambda}^{1/2} \quad \text{and} \quad \hat{\mathbf{F}} = \mathbf{V}\,\mathbf{\Lambda}^{1/2}$$

Equations 9.43b and 9.44b cannot be used here to find $\mathbf{F}$ and $\hat{\mathbf{F}}$ because the eigenanalysis has been conducted on a covariance matrix (eq. 11.16) computed from

the matrix of fitted values $\hat{\mathbf{Y}}$ (eq. 11.15) and not from **Q**. The site scores which are linear combinations of the environmental variables, corresponding to eq. 11.13 of RDA, are found from $\hat{\mathbf{Y}}$ using the following equations:

For scaling type 1: $$\mathbf{Z}_{\text{scaling 1}} = \mathbf{D}(p_{i+})^{-1/2}\,\hat{\mathbf{Y}}\,\mathbf{U} \qquad \textbf{(11.17)}$$

For scaling type 2: $$\mathbf{Z}_{\text{scaling 2}} = \mathbf{D}(p_{i+})^{-1/2}\,\hat{\mathbf{Y}}\,\mathbf{U}\grave{\mathbf{O}}^{-1/2} \qquad \textbf{(11.18)}$$

With scaling type 1, biplots can be drawn using either **F** and **V**, or $\mathbf{Z}_{\text{scaling 1}}$ and **V**. With scaling type 2, one can use either $\hat{\mathbf{V}}$ and $\hat{\mathbf{F}}$, or $\mathbf{Z}_{\text{scaling 2}}$ and $\hat{\mathbf{F}}$. The construction and interpretation of CCA biplots is discussed by ter Braak & Verdonschot (1995).

• Residuals can be analysed by applying eigenvalue decomposition (eq. 11.10) to matrix $\mathbf{Y}_{\mathbf{res}}$, producing matrices of eigenvalues $\grave{\mathbf{O}}$ and normalized eigenvectors **U**. Matrix $\hat{\mathbf{U}}$ is obtained using eq. 9.38: $\hat{\mathbf{U}} = \mathbf{Q}\,\mathbf{U}\grave{\mathbf{O}}^{-1/2}$. Species and site scores are obtained for scaling types 1 and 2 (eqs. 9.41, 9.42, 9.43a and 9.44a) using the matrices of row and column sums $\mathbf{D}(p_{i+})^{-1/2}$ and $\mathbf{D}(p_{+j})^{-1/2}$ of the original matrix **Y**.

A little-known application of CCA is worth mentioning here. Consider a qualitative environmental variable and a table of species presence-absence or abundance data. How can one "quantify" the qualitative states, i.e. give them values along a quantitative scale which would be related in some optimal way to the species data? CCA provides an easy answer to this problem. The species data form matrix **Y**; the qualitative variable, recoded as a set of dummy variables, is placed in matrix **X**. Compute CCA and take the fitted site scores ("site scores which are linear combinations of environmental variables"): they provide a quantitative rescaling of the qualitative variable, maximizing the weighted linear correlation between the dummy variables and matrix $\mathbf{Q}$. In the same way, RDA may be used to rescale a qualitative variable with respect to a table of quantitative variables of the objects if linear relationships can be assumed.

McCune (1997) warns users of CCA against inclusion of noisy or irrelevant explanatory variables in the analysis. They may lead to misleading interpretations.

2 — Numerical example

Table 11.3 will now be used to illustrate the computation and interpretation of CCA. The 9 species are used as matrix **Y**. Matrix **X** is the same as in Subsection 11.1.2. Results are presented in Table 11.5 and Fig. 11.5; programs such as CANOCO provide more output tables than presented here. There was a possibility of 3 canonical and 8 non-canonical axes. Actually, the last 2 non-canonical axes have zero variance. An overall test of significance (Subsection 11.3.2) showed that the canonical relationship between matrices **X** and **Y** is very highly significant ($p = 0.001$ after 999 permutations, by permutation of residuals under a full model; Subsection 11.3.2). The canonical axes explain 47%, 24% and 10% of the response table's variance, respectively. They are all significant ($p < 0.05$) and display strong row-weighted species-environment correlations (r = 0.998, 0.940, and 0.883, respectively).

Table 11.5 Results of canonical correspondence analysis of the data in Table 11.3 (selected output). Matrix **Y**: species 1 to 9; **X**: depth and 3 substrate classes. Non-canonical axes VIII and IX not shown.

	Canonical axes			Non-canonical axes			
	I	II	III	IV	V	VI	VII
Eigenvalues (their sum is equal to the total inertia in matrix $\bar{\mathbf{Q}}$ of species data = 0.78417)							
	0.36614	0.18689	0.07885	0.08229	0.03513	0.02333	0.00990
Fraction of the total variance in $\bar{\mathbf{Q}}$							
	0.46691	0.23833	0.10055	0.10494	0.04481	0.02975	0.01263
Cumulative fraction of total inertia in $\bar{\mathbf{Q}}$ accounted for by axes 1 to k							
	0.46691	0.70524	0.80579	0.91072	0.95553	0.98527	0.99791
Eigenvectors ("species scores"): matrices $\hat{\mathbf{F}}$ for the canonical and the non-canonical portions (eq. 9.44a)							
Species 1	–0.11035	–0.28240	–0.20303	0.00192	0.08223	0.08573	–0.01220
Species 2	–0.14136	–0.30350	0.39544	0.14127	0.02689	0.14325	0.04303
Species 3	1.01552	–0.09583	–0.19826	0.10480	–0.13003	0.02441	0.04647
Species 4	1.03621	–0.10962	0.22098	–0.22364	0.24375	–0.02591	–0.05341
Species 5	–1.05372	–0.53718	–0.43808	–0.22348	0.32395	0.12464	–0.11928
Species 6	–0.99856	–0.57396	0.67992	0.38996	–0.29908	0.32845	0.21216
Species 7	–0.25525	0.17817	–0.20413	–0.43340	–0.07071	–0.18817	0.12691
Species 8	–0.14656	0.85736	–0.01525	–0.05276	–0.35448	–0.04168	–0.19901
Species 9	–0.41371	0.70795	0.21570	0.69031	0.14843	–0.33425	–0.00629
Site scores ("sample scores"): matrices $\hat{\mathbf{V}}$ for the canonical and the non-canonical portions (eq. 9.42)							
Site 1	–0.71059	3.08167	0.21965	1.24529	1.07293	–0.50625	0.24413
Site 2	–0.58477	3.00669	–0.94745	–2.69965	–2.13682	0.81353	0.47153
Site 3	–0.76274	3.15258	2.13925	3.11628	2.30660	–0.69894	–1.39063
Site 4	–1.11231	–1.07151	–1.87528	–0.66637	1.10154	1.43517	–1.10620
Site 5	0.97912	0.06032	–0.69628	0.61265	–0.98301	0.31567	0.57411
Site 6	–1.04323	–0.45943	–0.63980	–0.28716	0.57393	–1.44981	1.70167
Site 7	0.95449	0.08470	0.13251	0.42143	0.11155	–0.39424	–0.67396
Site 8	–0.94727	0.10837	0.52611	0.00565	–1.26273	–1.06565	–1.46326
Site 9	1.14808	–0.49045	0.47835	–1.17016	1.00599	0.07350	0.08605
Site 10	–1.03291	–1.03505	2.74692	1.28084	–0.36299	1.98648	1.05356
Correlations of environmental variables with site scores							
Depth	0.18608	–0.60189	0.65814				
Coral	0.99233	–0.09189	–0.04614				
Sand	–0.21281	0.91759	0.03765				
Other subs.	–0.87958	–0.44413	0.02466				
Correlations of environmental variables with fitted site scores (for biplots)							
Depth	0.18636	–0.64026	0.74521				
Coral	0.99384	–0.09775	–0.05225				
Sand	–0.21313	0.97609	0.04263				
Other subs.	–0.88092	–0.47245	0.02792				
Centroids of sites with code "1" for BINARY environmental variables, in ordination diagram							
Coral	1.02265	–0.10059	–0.05376				
Sand	–0.66932	3.06532	0.13387				
Other subs.	–1.03049	–0.55267	0.03266				

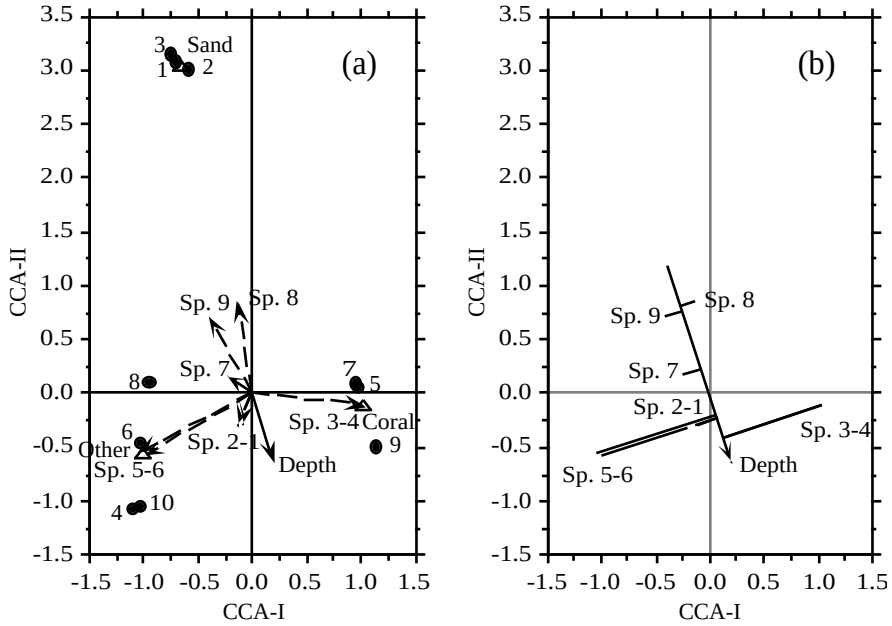


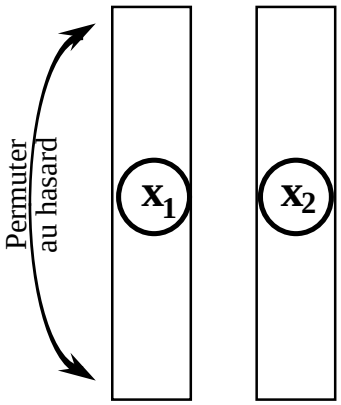
Figure 11.5 CCA ordination biplot of the artificial data in Table 11.3; the numerical results of the analysis are in Table 11.5. (a) Biplot representing the species (dashed arrows), sites (dots, with site identifiers which also correspond to depths in m) and environmental variables (full arrow for depth, triangles for the three binary substrate variables). (b) Ranking of the species along a quantitative environmental variable (depth in the present case) is inferred by projecting the species onto the arrow representing that variable.

Scaling type 2 (from Subsection 9.4.1) was used, in this example, to emphasize the relationships among species. As a result, the species (matrix $\hat{\mathbf{F}}$) are at the centroids of the sites (matrix $\hat{\mathbf{V}}$) in Fig. 11.5a and distances among species approximate their chi-square distances. Species 3 and 4 characterize the sites with coral substrate, whereas species 5 and 6 indicate the sites with "other substrate". Species 1 and 2, which occupy an intermediate position between the sites with coral and other substrate, are not well represented in the biplot of canonical axes I and II; axis III is needed to adequately represent the variance of these species. Among the ubiquitous species 7 to 9, two are well represented in the subspace of canonical axes I and II; their arrows fall near the middle of the area encompassing the three types of substrate. The sites are not perfectly ordered along the depth vector; the ordering of sites along this variable mainly reflects the difference in species composition between the shallow sandy sites (1, 2 and 3) and the other sites.

Tests de signification par permutations

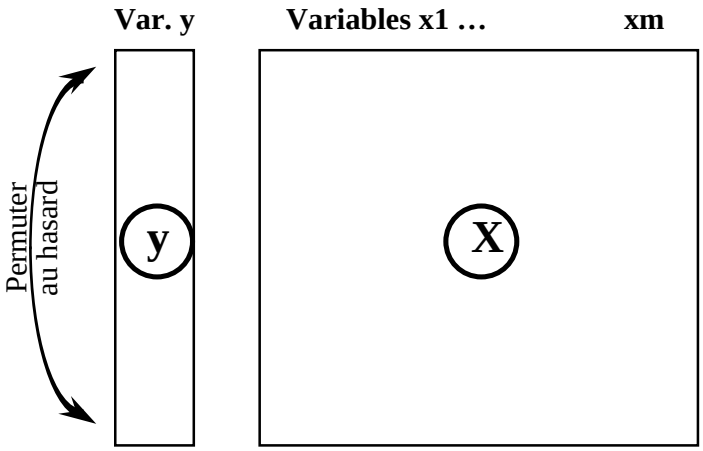
La distribution d'échantillonnage de la (ou des) statistique(s) est obtenue et permutant au hasard les lignes du premier vecteur ou de la première matrice et en recalculant la (ou les) statistique(s).
Le test s'effectue en comparant à cette (ces) distribution(s) d'échantillonnage la valeur de la (ou des) statistique(s) obtenue(s) pour les données non permutées.

I - Corrélation entre x_1 et x_2 :



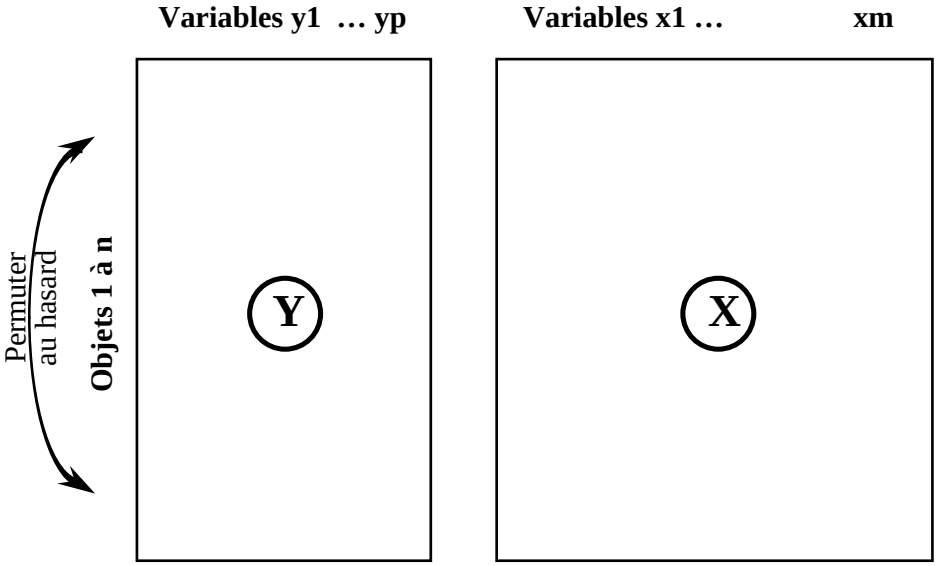
Statistique : *r* de Pearson ou
corrélation non-paramétrique

II - Régression multiple :



Statistiques : R^2
et coefficients de régression

III - Ordination de Y sous contrainte de X (analyse canonique) :



Statistiques :
- R^2 = proportion de la variation de **Y** expliquée par **X** = $\sum$ val. pr. can. $\Rightarrow F$
- Première valeur propre canonique

Sujets connexes

1. Transformation des tableaux de structure de communautés (présence-absence ou abondance d'espèces) pour les rendre utilisables en analyse canonique de redondance (ACR):

Legendre, P. and E. D. Gallagher. 2001. Ecologically meaningful transformations for ordination of species data. *Oecologia* 129: 271-280.

2. En présence de covariables, les tests par permutation en régression multiple et en analyse canonique se font par permutation des résidus d'un modèle nul ou d'un modèle complet, comme dans le programme Canoco. La théoorie de ces tests est expliquée dans l'article suivant:

Anderson, M. J. and P. Legendre. 1999. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *Journal of Statistical Computation and Simulation* 62: 271-303.

3.4- Using regression to carry out ANOVA

The ANOVA factor may be coded using dummy variables

Example, intercept + 3 groups: $\mathbf{b} = [\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{y}$

$$\mathbf{y} = \begin{bmatrix} y_{11} \\ \bullet \\ \bullet \\ y_{1j} \\ \hline y_{21} \\ \bullet \\ \bullet \\ y_{2j} \\ \hline y_{31} \\ \bullet \\ \bullet \\ y_{3j} \end{bmatrix} \qquad \mathbf{X} = \begin{array}{c} \begin{matrix} \text{Int.} & \multicolumn{3}{c}{\text{3 groups}} \end{matrix} \\ \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \hline 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ \hline 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \end{array} \qquad \mathbf{b} = \begin{bmatrix} \mu \\ b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

Example using 4 groups:

Groups	Dummy variables			
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1

Orthogonal dummy variables

2 groups: 1 variable	3 groups: 2 variables	4 groups: 3 variables	5 groups: 4 variables	etc.
$\begin{bmatrix} +1 \\ -1 \end{bmatrix}$	$\begin{bmatrix} +2 & 0 \\ -1 & +1 \\ -1 & -1 \end{bmatrix}$	$\begin{bmatrix} +3 & 0 & 0 \\ -1 & +2 & 0 \\ -1 & -1 & +1 \\ -1 & -1 & -1 \end{bmatrix}$	$\begin{bmatrix} +4 & 0 & 0 & 0 \\ -1 & +3 & 0 & 0 \\ -1 & -1 & +2 & 0 \\ -1 & -1 & -1 & +1 \\ -1 & -1 & -1 & -1 \end{bmatrix}$	

The number of (binary or orthogonal) dummy variables used in the regression equation is equal to the number of degrees of freedom of the among-group variation in ANOVA (= number of groups – 1).

Interaction

A significant interaction between two factors indicates that the effects of one of the factors are not consistent across the levels of the other factor.

	A_1	A_2
B_1	$\mu_{A_1B_1}$	$\mu_{A_2B_1}$
B_2	$\mu_{A_1B_2}$	$\mu_{A_2B_2}$

Hypothesis of no interaction

$$H_0: \mu_{A_1B_1} - \mu_{A_2B_1} = \mu_{A_1B_2} - \mu_{A_2B_2}$$

$$\mu_{A_1B_1} - \mu_{A_1B_2} = \mu_{A_2B_1} - \mu_{A_2B_2}$$

With more than one factor, one must represent the main factors using orthogonal dummy variables. Interaction dummy variables are the direct products of the dummy variables coding for the main factors. As a consequence, the dummy variables representing interactions will be linearly independent of the dummy variables representing the main factors if the main factors are crossed (two-way design) or nested.

Example: coding an interaction
Factor *A*: 3 groups ($a = 2$ orthogonal dummy variables)
Factor *B*: 2 groups ($b = 1$ dummy variable)
Interaction *AB*: ($a \times b$) dummy variables

$$\mathbf{y} = \begin{bmatrix} y_{111} \\ y_{112} \\ y_{113} \\ \hline y_{121} \\ y_{122} \\ y_{123} \\ \hline y_{211} \\ y_{212} \\ y_{213} \\ \hline y_{221} \\ y_{222} \\ y_{223} \\ \hline y_{311} \\ y_{312} \\ y_{313} \\ \hline y_{321} \\ y_{322} \\ y_{323} \end{bmatrix} \quad \mathbf{X} = \begin{array}{c} \\ \\ \\ \hline \\ \\ \\ \hline \\ \\ \\ \hline \\ \\ \\ \hline \\ \\ \\ \hline \\ \\ \end{array} \begin{array}{c} \overline{\quad A \quad} \\ \\ \\ \hline \\ \\ \\ \hline \\ \\ \\ \hline \\ \\ \\ \hline \\ \\ \\ \hline \\ \\ \end{array} \begin{array}{c} \\ \\ \\ \hline \\ \\ \\ \hline \\ \\ \\ \hline \\ \\ \\ \hline \\ \\ \\ \hline \\ \\ \end{array} \begin{array}{c} B \\ \\ \\ \hline \\ \\ \\ \hline \\ \\ \\ \hline \\ \\ \\ \hline \\ \\ \\ \hline \\ \\ \end{array} \begin{array}{c} \overline{\quad AB \quad} \\ \\ \\ \hline \\ \\ \\ \hline \\ \\ \\ \hline \\ \\ \\ \hline \\ \\ \\ \hline \\ \\ \end{array} \begin{bmatrix} +2 & 0 & 1 & +2 & 0 \\ +2 & 0 & 1 & +2 & 0 \\ +2 & 0 & 1 & +2 & 0 \\ \hline +2 & 0 & -1 & -2 & 0 \\ +2 & 0 & -1 & -2 & 0 \\ +2 & 0 & -1 & -2 & 0 \\ \hline -1 & 1 & 1 & -1 & 1 \\ -1 & 1 & 1 & -1 & 1 \\ -1 & 1 & 1 & -1 & 1 \\ \hline -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 \\ \hline -1 & -1 & 1 & -1 & -1 \\ -1 & -1 & 1 & -1 & -1 \\ -1 & -1 & 1 & -1 & -1 \\ \hline -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 & 1 \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} b_{A1} \\ b_{A2} \\ \hline b_{B1} \\ \hline b_{AB11} \\ b_{AB21} \end{bmatrix}$$

4- Example 1: ANOVA with 2 crossed factors, computed using RDA

Example 1: Data from Sokal & Rohlf (1981, p. 325). The data may be recoded as follows for RDA:

$$\text{Food consumption} = \begin{bmatrix} 709 \\ 679 \\ 699 \\ 592 \\ 538 \\ 476 \\ 657 \\ 594 \\ 677 \\ 508 \\ 505 \\ 539 \end{bmatrix} \quad \text{Sex} = \begin{bmatrix} +1 \\ +1 \\ +1 \\ +1 \\ +1 \\ +1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \end{bmatrix} \quad \text{Lard} = \begin{bmatrix} +1 \\ +1 \\ +1 \\ -1 \\ -1 \\ -1 \\ +1 \\ +1 \\ +1 \\ -1 \\ -1 \\ -1 \end{bmatrix} \quad \text{Interaction} = \begin{bmatrix} +1 \\ +1 \\ +1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ +1 \\ +1 \\ +1 \end{bmatrix}$$

Two-way analysis of variance computed using RDA (n = 12):

	d.f.	RDA (CANOCO, 999 permutations)				Sokal & Rohlf	
		∑ can. λ	∑ can. λ/d.f.	*F*	Prob.	*F*	Prob.
Sex	1	0.0487	0.0487	2.593	0.140	2.593	0.1460
Lard	1	0.7890	0.7890	41.969	0.001	41.969	0.0002
Interact.	1	0.0118	0.0118	0.630	0.463	0.630	0.4503
Resid.	8	0.1504	0.0188				

"∑ can. λ" for residuals = sum of the non-canonical eigenvalues. This value is given in each of the three CANOCO output files which are needed to obtain the "∑ can. λ" for Sex, Lard and Interaction.

Analyse de variance par analyse de redondance (ACR)

Le codage en variables orthogonales (contrastes de Helmert[1]) est expliqué en détail à l'annexe C de l'article suivant:

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

Exemple d'analyse de variance multivariable d'un tableau d'abondances d'espèces (pages suivantes):

Hooper, E., R. Condit and P. Legendre. 2002. Responses of 20 native tree species to reforestation strategies for abandoned farmland in Panama. *Ecological Applications* 12(6): 1626-1641.

[1] Voir pages 155-156 de:
Venables, W. N. and B. D. Ripley. 1994. *Modern applied statistics with S-Plus*. Springer-Verlag, New York.

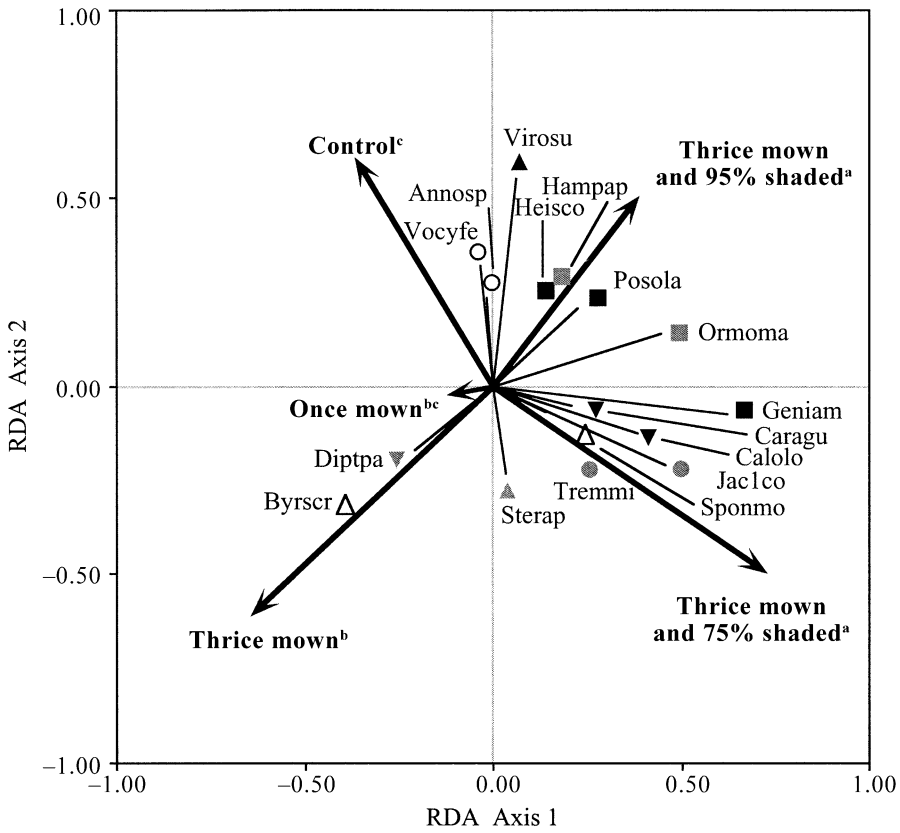


FIG. 6. Ordination biplot illustrating the significant ($P < 0.001$) effect of *Saccharum* treatment on the relative height growth (RHT) of 16 tree species in the wet season, between May and July 1997. Same-letter superscripts indicate no significant difference ($P < 0.05$) based on post hoc analysis. *Saccharum* treatment explained 19.8% of the variance between species, with the first and second axes explaining 11.2% and 4.5%, respectively. Arrows indicate treatments and lines indicate species vectors. Codes associated with each species are listed in Table 1, and the symbols match Fig. 2.

relatively high performance in the *Saccharum* control (Table 4). We conclude that the *Saccharum* did not completely limit their regeneration. A reforestation effort starting with seed and minimal pre-sowing treatment is likely to succeed with these large-seeded, shade-tolerant species.

Fires burn yearly in the dry season in these *Saccharum*-dominated grasslands, and our data show that these wildfires are also a major barrier to tree regeneration. Fire killed most species and significantly lowered the germination of all except *Trema* and *Byrsonima* (but those two cannot compete with established *Saccharum*). Resprouting from cut stems or stumps is a very common mechanism for reestablishment following disturbance (Aide et al. 1995), and we found that seedlings of several large-seeded species (*Carapa*, *Dipteryx*, *Virola*, *Ormosia*, and *Calophyllum*) could resprout following fire. Recurring fires, as a result of grass invasion following pasture abandonment, arrest natural tree regeneration in abandoned pastures at other Neotropical sites as well (Janzen 1988, Nepstad et al. 1990, Aide and Cavelier 1994).

Management suggestions

Fire is a major barrier to tree regeneration at our sites, limiting both establishment and species diversity. We therefore recommend the establishment of firebreaks, which have also been an integral part of reforestation strategies in Costa Rica (Janzen 1988) and the Amazon (Nepstad et al. 1990). The breaks must be large for effective fire protection because the flame height of *Saccharum* wildfires can reach >15 m.

Many alternatives have been suggested for forest restoration throughout the wet tropics. These range, in order of increasing cost, from simply allowing natural regeneration to proceed, to planting seeds or seedlings to assist natural regeneration, through establishing tree plantations and allowing recruitment of tree seedlings below them (Brown and Lugo 1994, Guariguata et al. 1995, Kuusipalo et al. 1995). Our goal was to find a low-cost strategy for extensive forest restoration in abandoned Panamanian farmland, and our results suggest that even with the removal of fire, natural tree regeneration will not proceed unassisted because the *Saccharum* poses a formidable barrier to the small-

APPENDIX A

Summary of repeated-measures ANOVA on tree seedling germination, abundance, survival, height, and relative height growth (RHT).

Source	Germination†		Abundance†		Survival†		Height‡		RHT‡	
	df	*F*	df	*F*	df	*F*	df	*F*	df	*F*
Between subjects										
Distance	2	0.52	2	3.39	2	1.11	2	3.52	2	4.77
Site × distance	4§									
Treatment	4	11.68**	4	13.04**	4	2.79*	4	3.22**	4	4.86**
Treatment × distance	8	0.69	8	0.45	8	0.43	8	0.22	8	0.58
Site × treatment + Site × treatment × distance	24‖									
Within subjects										
Time	2	36.31**	3	80.73**	2	1.71	2	5.68**	2	0.86
Time × distance	4	0.59	6	1.01	4	0.15	4	2.17	4	0.13
Time × site × distance	8§									
Time × treatment	8	1.84	12	2.82**	8	1.09	8	1.08	8	0.26
Time × treatment × distance	16	0.92	24	0.80	16	0.74	16	0.36	16	1.08
Time × site × treatment + Time × site × treatment × distance	48‖									

Notes: The analysis followed a repeated-measures, split-plot design. Sources of variation included distance from the forest as the main plot factor (distance), shading and mowing treatments of the *Saccharum* as the subplot factor (treatment), and their interactions. Site was included as a blocking factor.

* $P < 0.05$; ** $P < 0.01$.

† Sample size: $n = 45$ subjects (i.e., 45 unburned subplots).

‡ Sample size: $n = 371$ subjects (i.e., 371 seedlings were present over all three time periods in the 45 unburned subplots).

§ Main plot error.

‖ Subplot error.

APPENDIX B

Summary of RDA in a MANCOVA-like design on matrices of germination per species per subplot per time period (germination) or relative growth rates for height (RHT) per species per subplot per time period.

Source	Germination		RHT	
	df†	*F*	df‡	*F*
Site	2		2	
Distance	2	1.03	2	1.41
Treatment	4	2.89**	4	4.67**
Treatment × distance	8	0.70	8	1.25
Time	3	31.51**	1	12.08**
Time × distance	6	1.01	2	2.46*
Time × treatment	12	1.81**	4	2.85**
Time × treatment × distance	24	0.72	8	1.36

Notes: Sources of variation included distance from the forest (distance), shading and mowing treatments of the *Saccharum* (treatment), time, and their interactions. Site and the interaction of site with all main factors and interactions were used as covariables.

* $P < 0.05$; ** $P < 0.01$ (determined using permutation testing).

† For all factors and interactions, denominator df = 60.

‡ For all factors and interactions, denominator df = 16.

Étude de la variation des peuplements d'insectes dans des zones humides du Vénézuéla au cours du temps:

Grillet, M. E., P. Legendre and D. Borcard. 2002. Community structure of neotropical wetland insects in Northern Venezuela. I. Temporal and environmental factors. *Archiv für Hydrobiologie* 155(3): 413-436.

Les mois de l'année forment une variable qualitative multiclasse qui est l'équivalent d'un critère de classification en anova.

Des diagrammes de double projection (« *biplots* ») montrent les relations entre les familles d'insectes et les mois de l'année, mettant en évidence les différences entre saison des pluies et saison sèche.

Quelles sont les échelles spatiales importantes dans un écosystème?

Pierre Legendre

Département de sciences biologiques

Université de Montréal

Pierre.Legendre@umontreal.ca

http://www.bio.umontreal.ca/legendre/

Modélisation spatiale dans le cadre des sciences géomatiques, géographiques et forestières

Colloque de l'AGÉOFOR, Université Laval, Québec, 15 avril 2005

Plan de l'exposé

1. Introduction
2. Partition de la variation
3. Analyse multi-échelle
   - La méthode CPMV
   - Simulations numériques
4. Application à des données écologiques réelles

Cadre conceptuel

Les écologistes cherchent à comprendre l'organisation spatio-temporelle des communautés par l'étude des assemblages d'espèces.

Les assemblages d'espèces forment la meilleure variable réponse pour estimer l'impact des changements que subissent les écosystèmes.

Difficulté: assemblages = tableaux multivariables (sites x espèces).

La présence de **structures spatiales** dans les communautés indique l'existence d'un processus générateur. Deux types de mécanismes peuvent générer des structures spatiales dans les communautés:

• l'autocorrelation de l'assemblage d'espèces (variables réponse);

• le forçage par des variables explicatives: contrôle environnemental ou biotique des assemblages, ou encore dynamique historique.

Pour arriver à comprendre/modéliser les mécanismes qui génèrent ces structures, il nous faut incorporer explicitement les structures spatiales, à toutes les échelles, dans le modèle statistique.

Partitionner la variation multivariable

Borcard, Legendre & Drapeau 1992 (422 citations)

Borcard & Legendre 1994

De nombreux articles publiés utilisent cette méthode

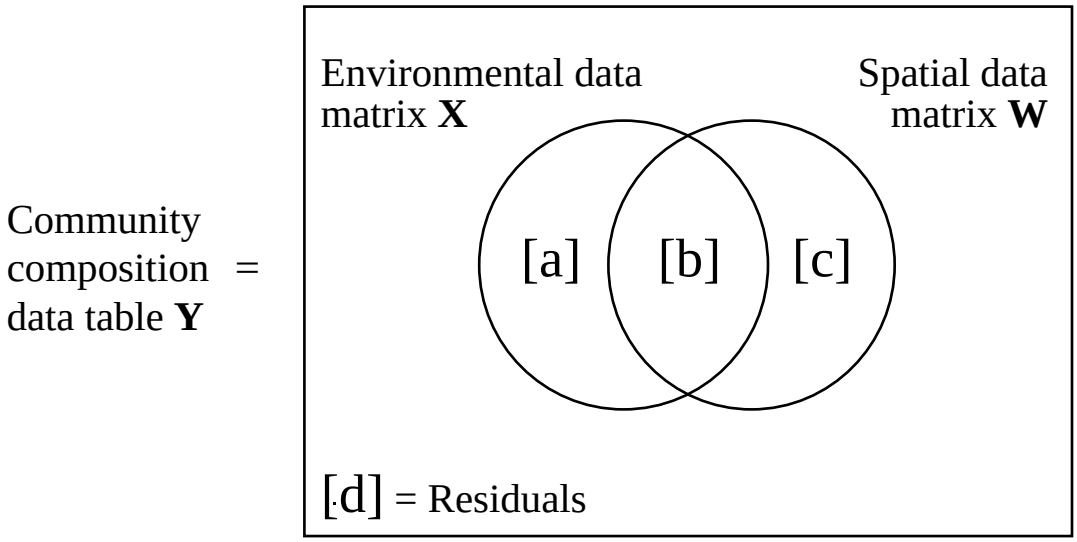
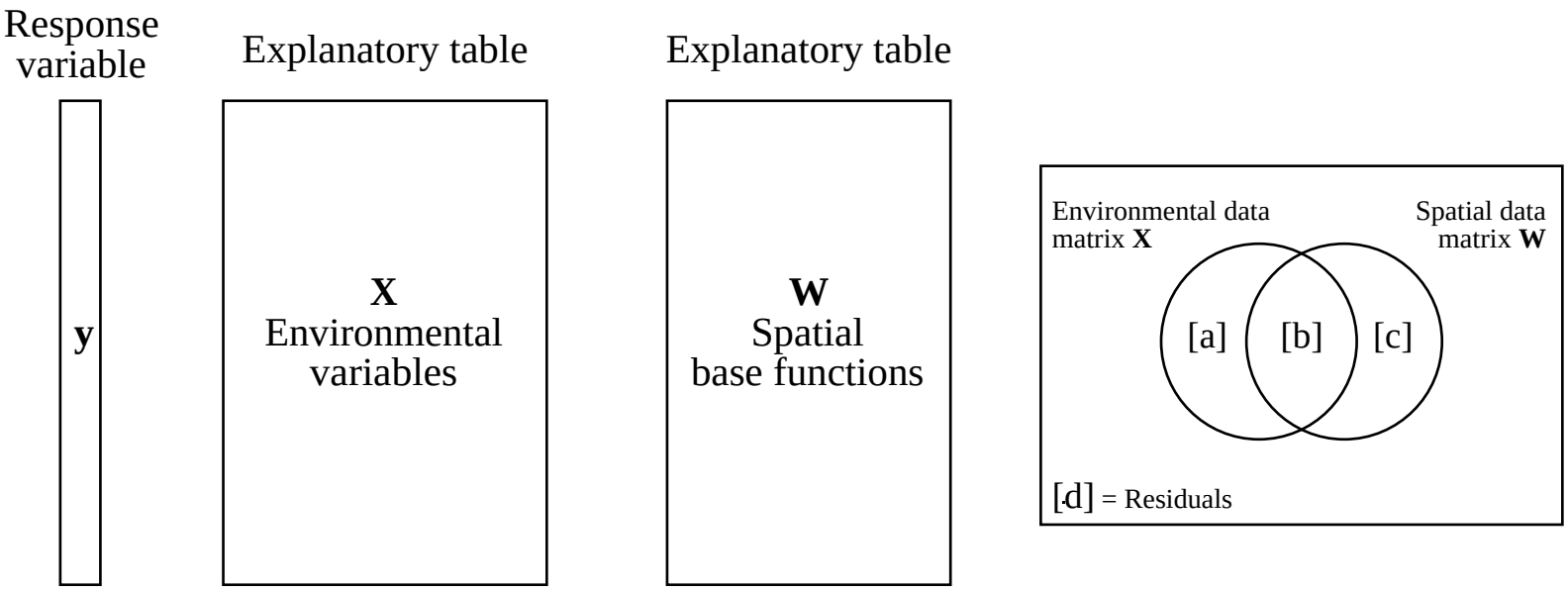


Figure – Diagramme de Venn illustrant la partition de la variation d'une matrice **Y** (par exemple, sites x espèces) à l'aide de tableaux explicatifs de variables environnementales (matrice **X**) et spatiales (matrice **W**). Le rectangle représente 100% de la variation de **Y**.

Méthode décrite dans les articles et textes suivants :
Borcard, Legendre & Drapeau 1992, Borcard & Legendre 1994, Legendre & Legendre 1998.

Comment arrive-t-on à combiner des variables environnementales et spatiales dans un modèle statistique?

• Cas #1 – une seule variable réponse **y** : régression partielle[1].



[1] Méthode décrite dans Legendre & Legendre (1998). Nouvelle rédaction de la section 10.3.5 "Partial linear regression and variation partitioning" disponible parmi les "Documents divers" de la page http://biol10.biol.umontreal.ca/BIO6077/.

Comment arrive-t-on à combiner des variables environnementales et spatiales dans un modèle statistique?

• Cas #2 – données multivariables **Y** : analyse canonique partielle[1] (ACR ou ACC).

Response table

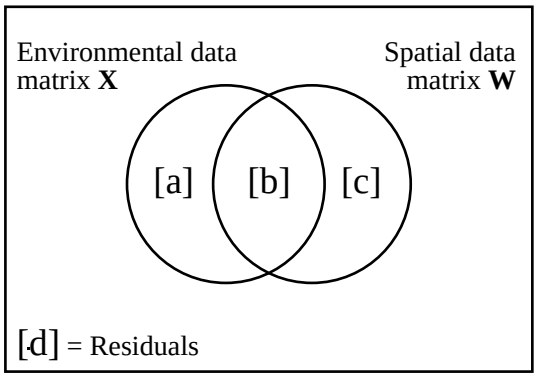
Y
Community composition data

Explanatory table

X
Environmental variables

Explanatory table

W
Spatial base functions



[1] Méthode décrite dans Legendre & Legendre (1998, chapitre 13).

Fonctions géographiques

Première représentation (simple) : polynôme des coordonnées géographiques (analysis des surfaces polynomiales).

Exemple 1 : 20 points d'échantillonnage dans l'étang de Thau, sud de la France.

	X	Y	X centred	Y centred	X^2	XY	Y^2	X^3	X^2Y	XY^2	Y^3
1	3	10	-8.75	3.70	76.5625	-32.3750	13.6900	-669.92188	283.28125	-119.78750	50.65300
2	5	9	-6.75	2.70	45.5625	-18.2250	7.2900	-307.54688	123.01875	-49.20750	19.68300
3	6	8	-5.75	1.70	33.0625	-9.7750	2.8900	-190.10938	56.20625	-16.61750	4.91300
4	6	10	-5.75	3.70	33.0625	-21.2750	13.6900	-190.10938	122.33125	-78.71750	50.65300
5	8	9	-3.75	2.70	14.0625	-10.1250	7.2900	-52.73438	37.96875	-27.33750	19.68300
6	9	10	-2.75	3.70	7.5625	-10.1750	13.6900	-20.79688	27.98125	-37.64750	50.65300
7	10	7	-1.75	.70	3.0625	-1.2250	.4900	-5.35938	2.14375	-.85750	.34300
8	11	6	-.75	-.30	.5625	.2250	.0900	-.42188	-.16875	-.06750	-.02700
9	12	5	.25	-1.30	.0625	-.3250	1.6900	.01562	-.08125	.42250	-2.19700
10	12	7	.25	.70	.0625	.1750	.4900	.01562	.04375	.12250	.34300
11	12	9	.25	2.70	.0625	.6750	7.2900	.01562	.16875	1.82250	19.68300
12	13	8	1.25	1.70	1.5625	2.1250	2.8900	1.95312	2.65625	3.61250	4.91300
13	15	2	3.25	-4.30	10.5625	-13.9750	18.4900	34.32812	-45.41875	60.09251	-79.50701
14	15	4	3.25	-2.30	10.5625	-7.4750	5.2900	34.32812	-24.29375	17.19250	-12.16700
15	15	5	3.25	-1.30	10.5625	-4.2250	1.6900	34.32812	-13.73125	5.49250	-2.19700
16	16	1	4.25	-5.30	18.0625	-22.5250	28.0900	76.76562	-95.73125	119.38251	-148.87702
17	16	2	4.25	-4.30	18.0625	-18.2750	18.4900	76.76562	-77.66875	78.58251	-79.50701
18	16	4	4.25	-2.30	18.0625	-9.7750	5.2900	76.76562	-41.54375	22.48250	-12.16700
19	17	6	5.25	-.30	27.5625	-1.5750	.0900	144.70312	-8.26875	.47250	-.02700
20	18	4	6.25	-2.30	39.0625	-14.3750	5.2900	244.14062	-89.84375	33.06250	-12.16700

$$\hat{z} = f(X,Y) = b_0 + b_1X + b_2Y + b_3X^2 + b_4XY + b_5Y^2 + b_6X^3 + b_7X^2Y + b_8XY^2 + b_9Y^3$$

Petit exemple pédagogique

20 points d'échantillonnage dans l'étang de Thau, sud de la France.

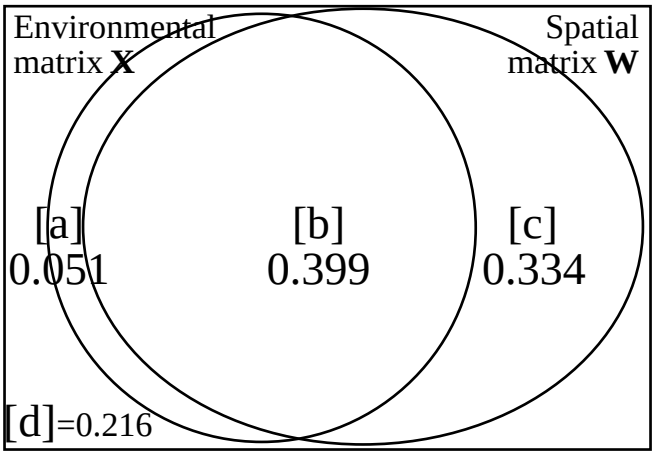
Variables

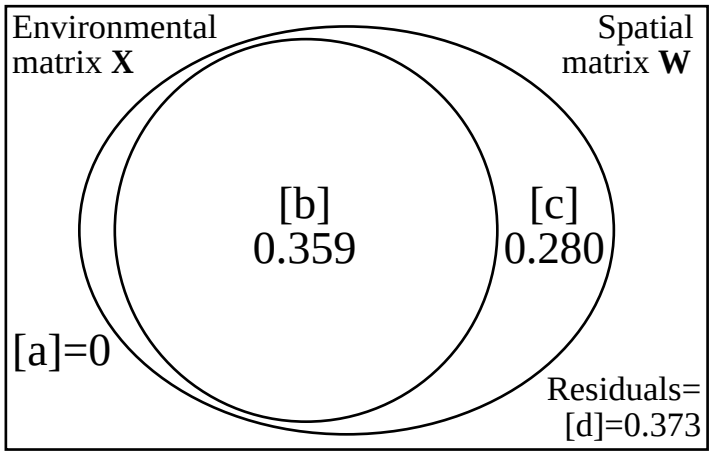
Réponse (**Y**) : 2 types de bactéries hétérotrophes aérobies (log)

Environmentales (**X**) : NH_4, phéopigments, production bactérienne

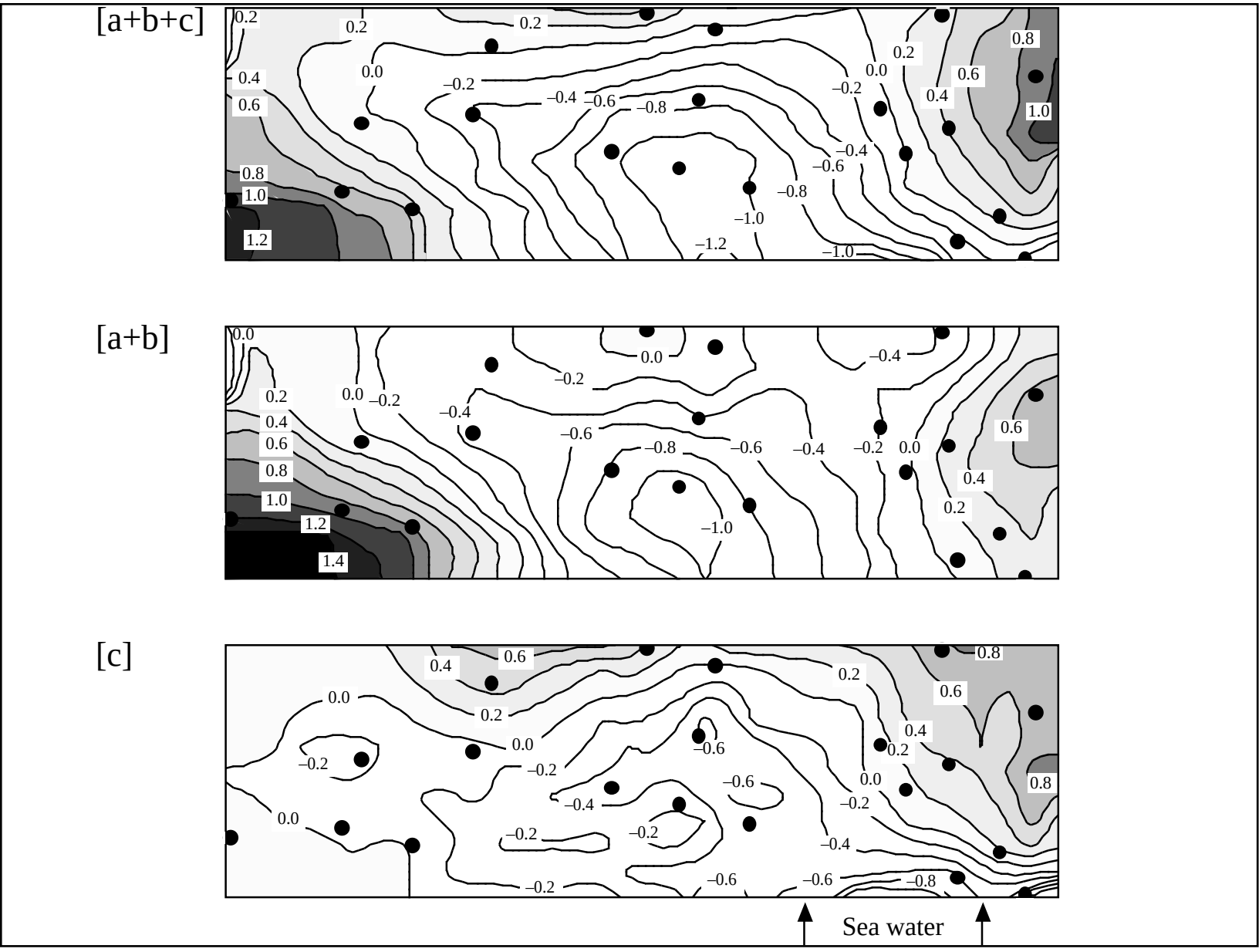
Spatiales (**W**) : monomes géographiques X^2, X^3, X^2Y, XY^2, Y^3

Dans les études réelles, on analyse des tableaux de données contenant beaucoup plus d'observations (n).





Fractions of variation	**Proportion of variation of Y (R^2)**	**Probability (999 perm.)**	**Adjusted R^2**
[a+b]	0.450	0.005*	0.347
[b+c]	0.734	0.001*	0.639
[a+b+c]	0.784	0.001*	0.627
[a]	0.051	0.549	$-0.012 \approx 0$
[b]	0.399	*Cannot be tested*	0.359
[c]	0.334	0.011*	0.280
Residuals = [d]	0.216		0.373
[a+b+c+d]	1.000		1.000


[a+b+c]
0.2
0.2
0.2
0.8
0.2
0.0
0.0
0.4
0.6
0.4
0.6
–0.2
–0.2
–0.4
–0.6
–0.8
–0.4
–0.6
–0.8
1.0
0.8
1.0
–1.0
–1.2
1.2
–1.0
[a+b]
0.0
0.0
–0.4
–0.2
0.0
–0.2
0.2
–0.4
0.4
–0.6
0.6
–0.8
–0.6
–0.4
–0.2
0.0
0.6
0.8
0.4
1.0
0.2
1.2
–1.0
1.4
[c]
0.4
0.6
0.8
0.2
0.6
0.0
0.2
0.4
0.0
–0.6
0.2
–0.2
–0.2
–0.6
0.0
0.8
–0.4
–0.2
0.0
–0.2
–0.2
–0.4
–0.2
–0.6
–0.6
–0.8
Sea water

Fonctions géographiques

Seconde représentation :

Coordonnées Principales de Matrices de Voisinage (CPMV)

(Principal Coordinates of Neighbour Matrices, CPMV)

Borcard & Legendre 2002, 2004, 2005

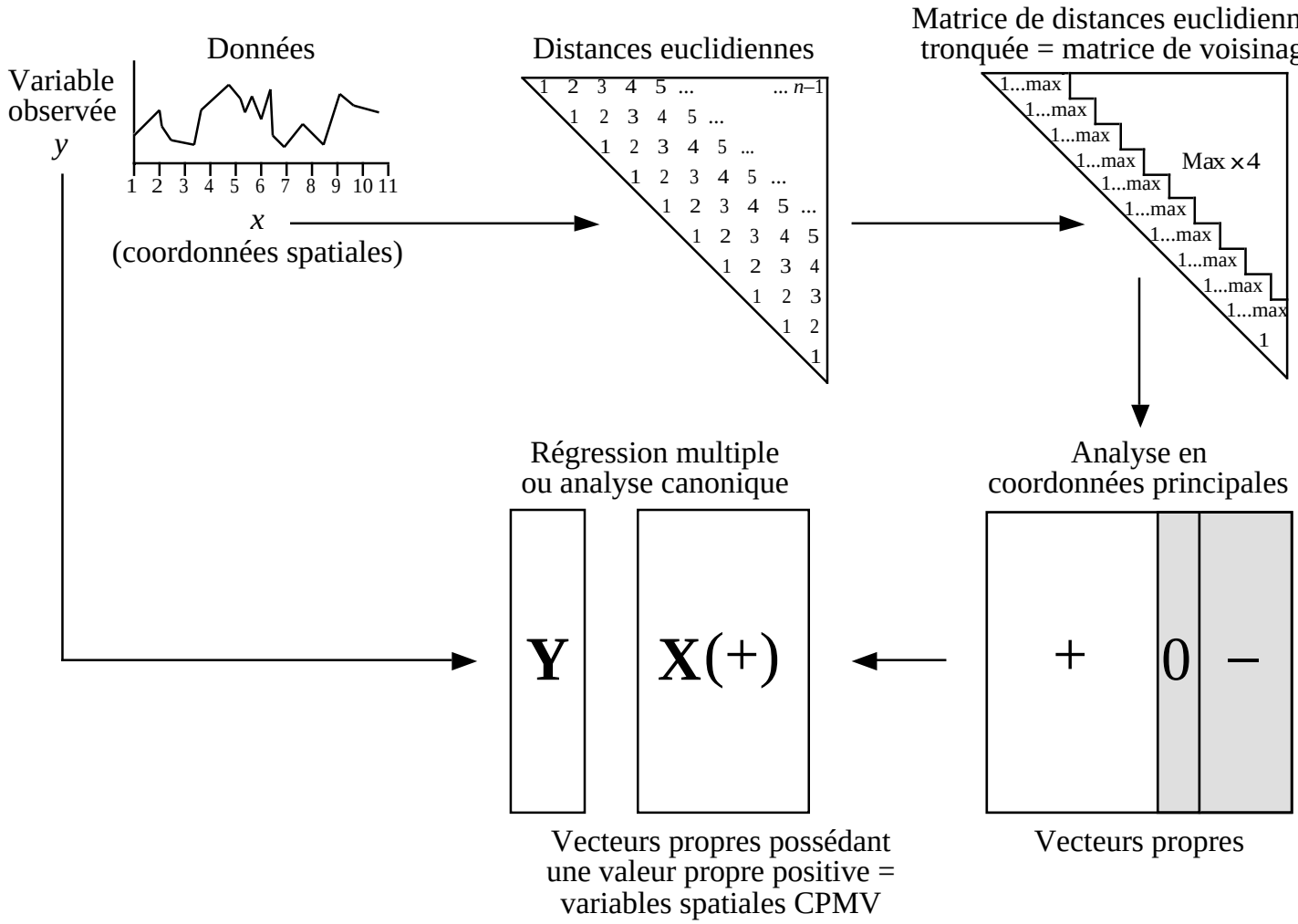


Figure – Description schématique de l'analyse CPMV. Les fonctions géographiques CPMV sont obtenues par analyse en coordonnées principales d'une matrice tronquée des distances géographiques entre les sites d'échantillonnage.

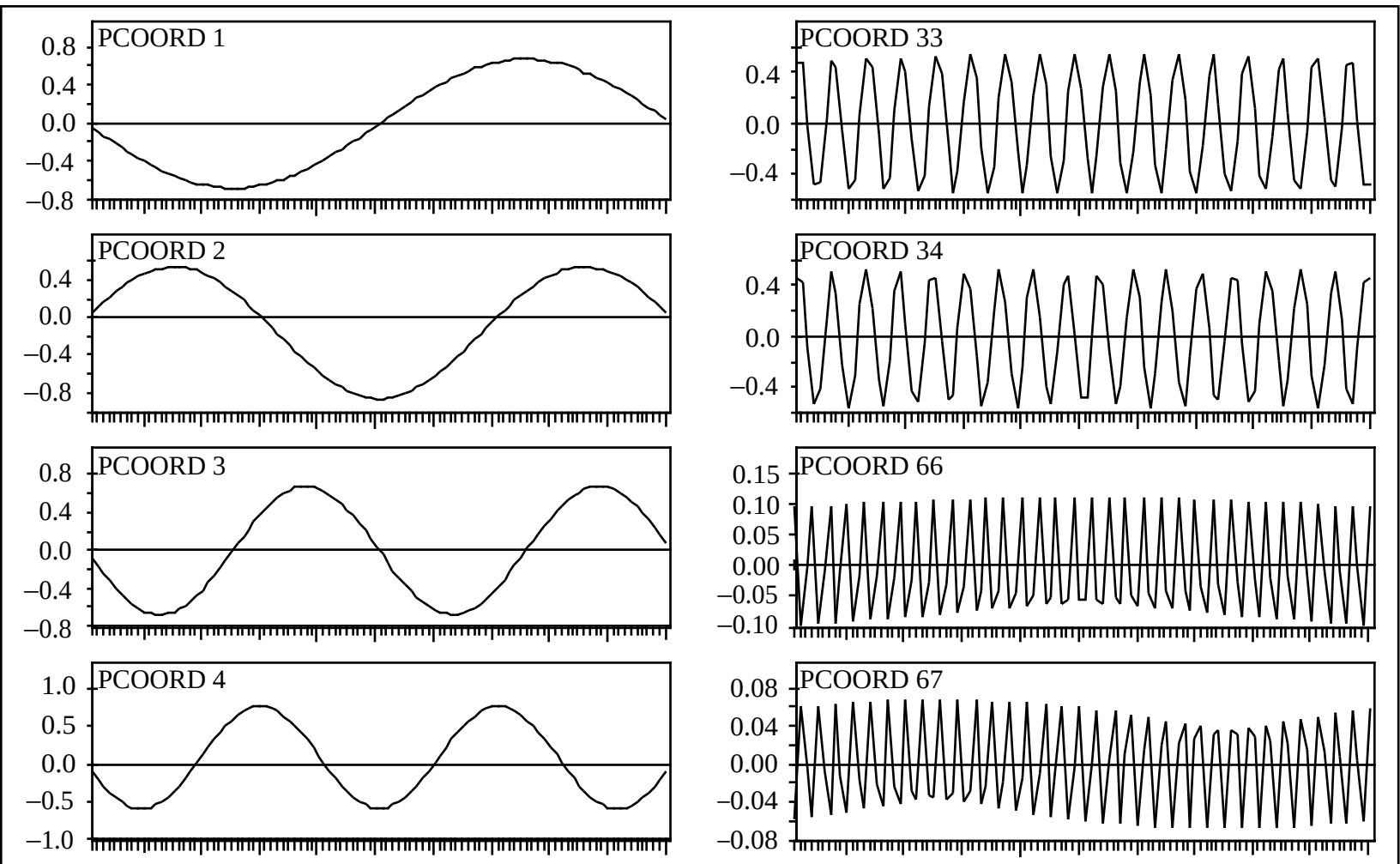


Figure – Huit des 67 fonctions géographiques CPMV obtenues pour un transect formé de 100 points équidistants. On a tronqué après le premier voisin.

Notes sur les fonctions géographiques CPMV

Les variables CPMV représentent une **décomposition spectrale** des relations spatiales entre les sites d'échantillonnage. On peut les calculer pour des points à espacement régulier ou irrégulier, dans l'espace ou le temps.

Les fonctions géographiques CPMV sont **orthogonales** (produit scalaire = 0). Si le pas d'échantillonnage est régulier, elles ont la forme de sinusoïdes. Il s'agit là d'une propriété de la décomposition en valeurs et vecteurs propres d'une matrice de distances centrée par lignes et par colonnes (Laplacien).

Simulations numériques

Étude de l'erreur de type I

Les simulations ont montré que l'analyse est honnête. En l'absence de structures spatiales, la procédure ne génère pas de résultats significatifs au-delà de ce qui est prévu par le niveau de signification α du test.

Étude de la puissance

Les simulations ont montré que l'analyse CPMV est capable de détecter des structures spatiales de différents types :

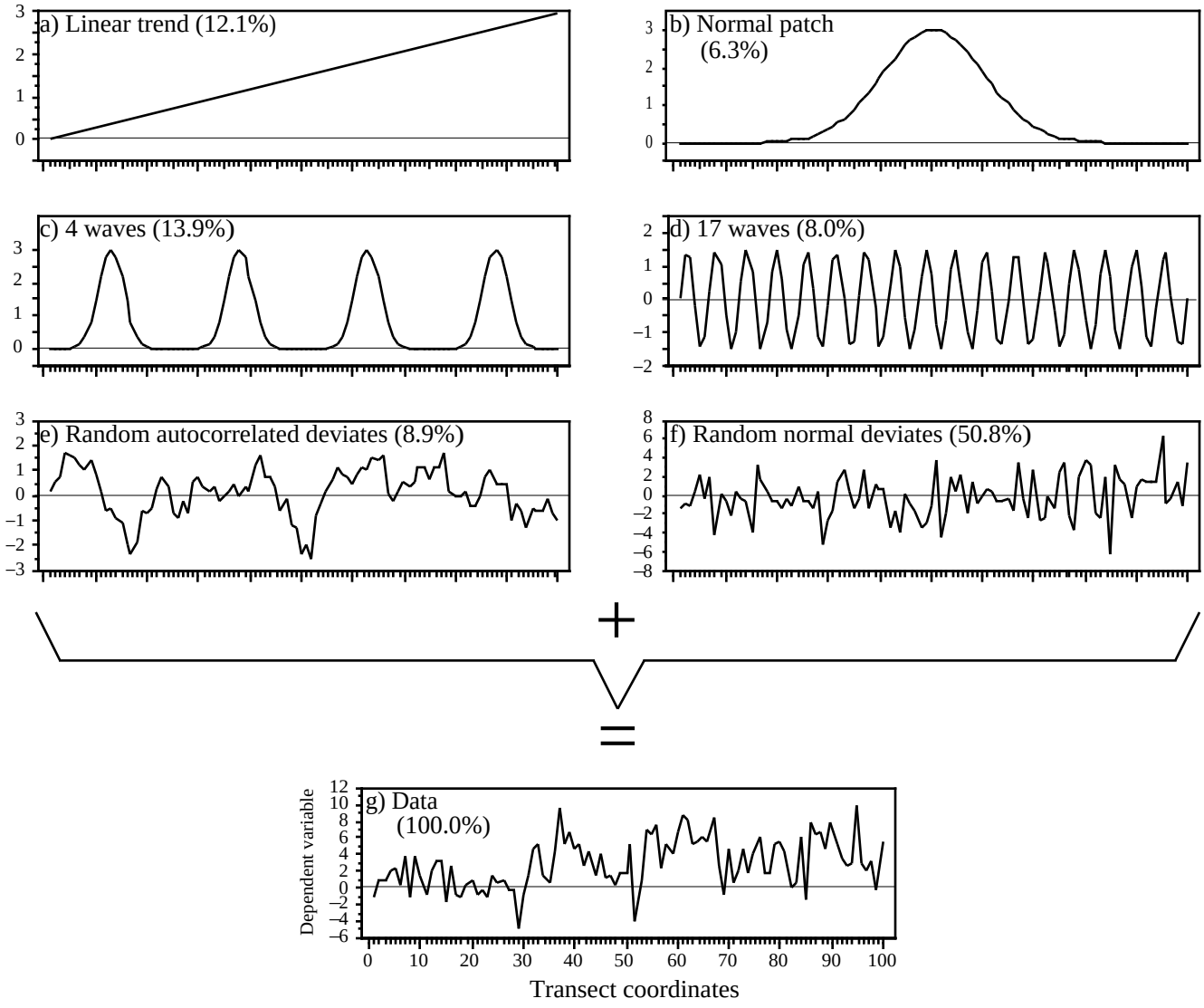
• données aléatoires autocorrélées,

• bosses et fonctions sinus de différentes tailles, avec ou sans bruit aléatoire, représentant des structures déterministes,

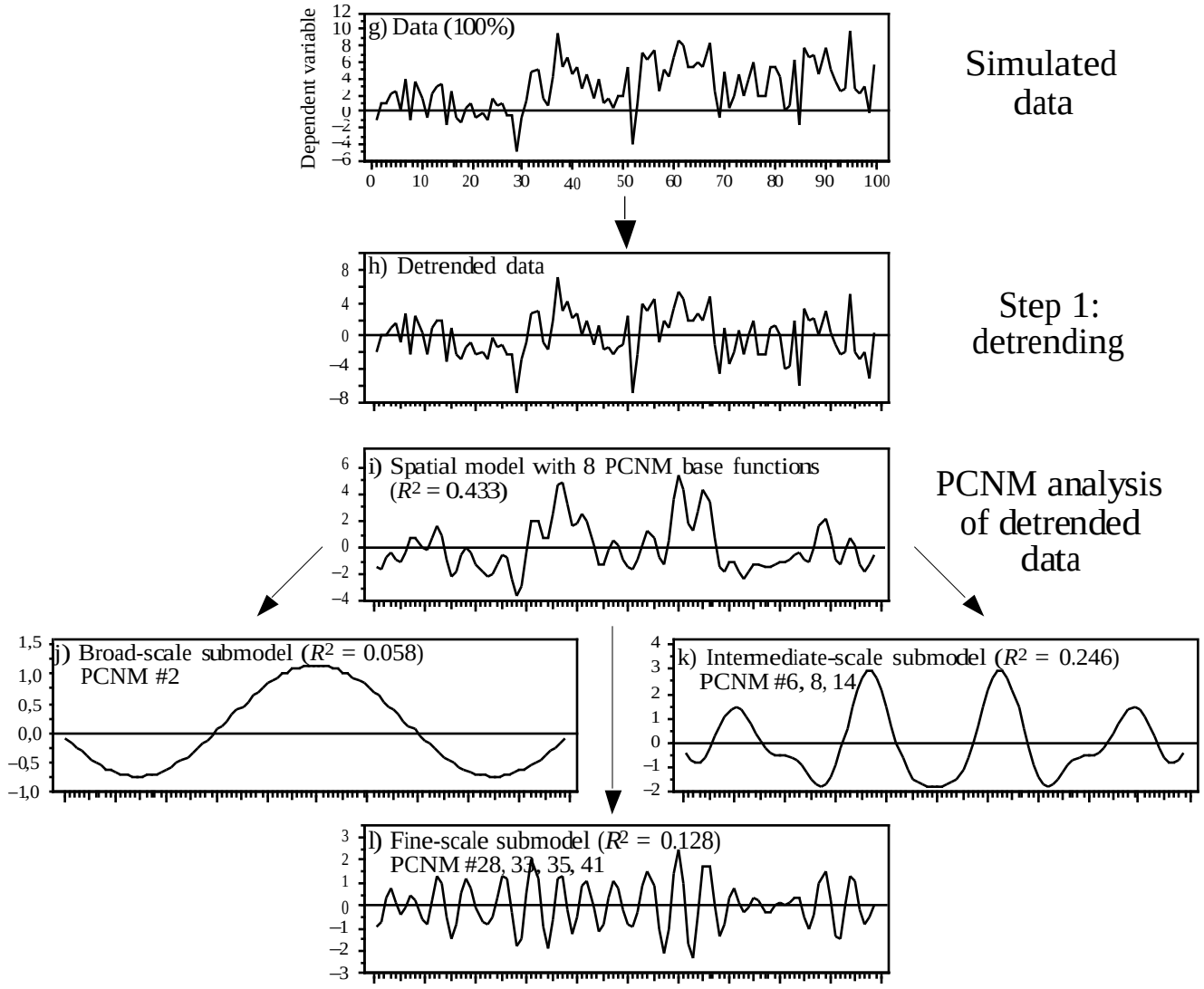
$\Rightarrow$ pour autant que les structures à détecter soient plus grandes que le niveau de tronquage utilisé pour créer les fonctions CPMV.

Résultats détaillés présentés dans Borcard & Legendre 2002.

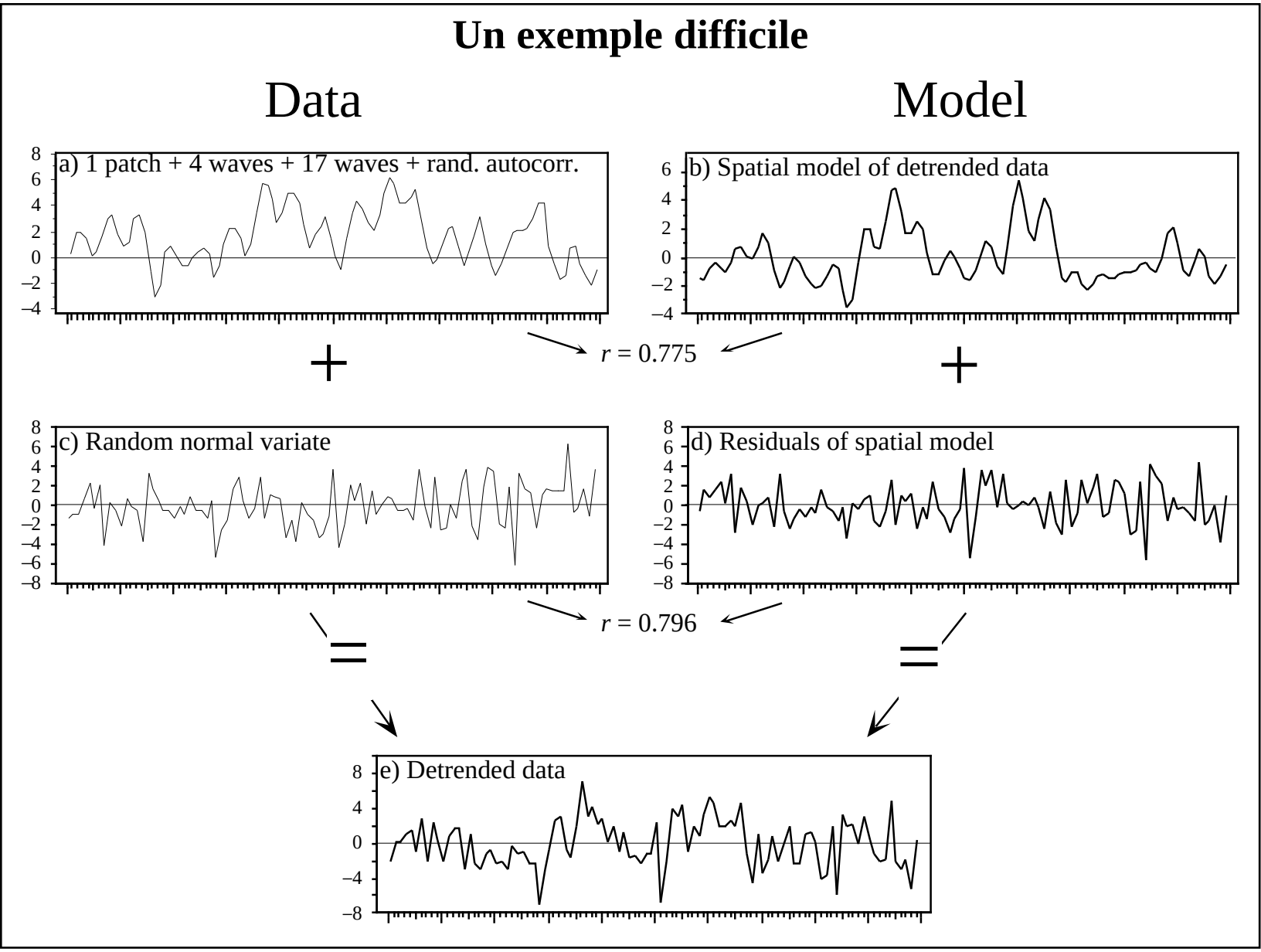
Un exemple difficile



Un exemple difficile



Un exemple difficile



Sélection des variables explicatives

Test de signification d'un modèle CPMV : on utilise toutes les variables CPMV sans sélection.

Le *coefficient de redondance bimultivariable ajusté* (R^2 ajusté), qui contient une correction pour le nombre de variables explicatives, produit une estimation non biaisée de la variation de **Y** expliquée par les fonctions CPMV.

Avant de représenter les fractions de variation dans des *diagrammes de double projection* (« *biplots* »), on procède habituellement à une sélection pas à pas des variables explicatives. On obtient ainsi des modèles parcimonieux à représenter dans les diagrammes. On sélectionne séparément les variables environnementales et les fonctions géographiques CPMV.

Exemple 1

Transect régulier en Amazonie péruvienne[1]

Données : abondance de la fougère *Adiantum tomentosum* dans des quadrats.

Plan d'échantillonnage : 260 quadrats adjacents (5 m x 5 m) formant un transect dans la region de Nauta au Pérou.

Questions

• À quelles échelles spatiales ces abondances sont-elle structurée ?

• Ces échelles sont-elles reliées à celles des var. environnementales ?

Pré-traitement

• Abondances de *Adiantum tomentosum* : racine carrée.

• Détendançage spatial (tendance linéaire : $R^2 = 0.102$, p = 0.001).

[1] Données de Tuomisto & Poulsen 2000, ré-analysées par Borcard, Legendre, Avois-Jacquet & Tuomisto 2004.

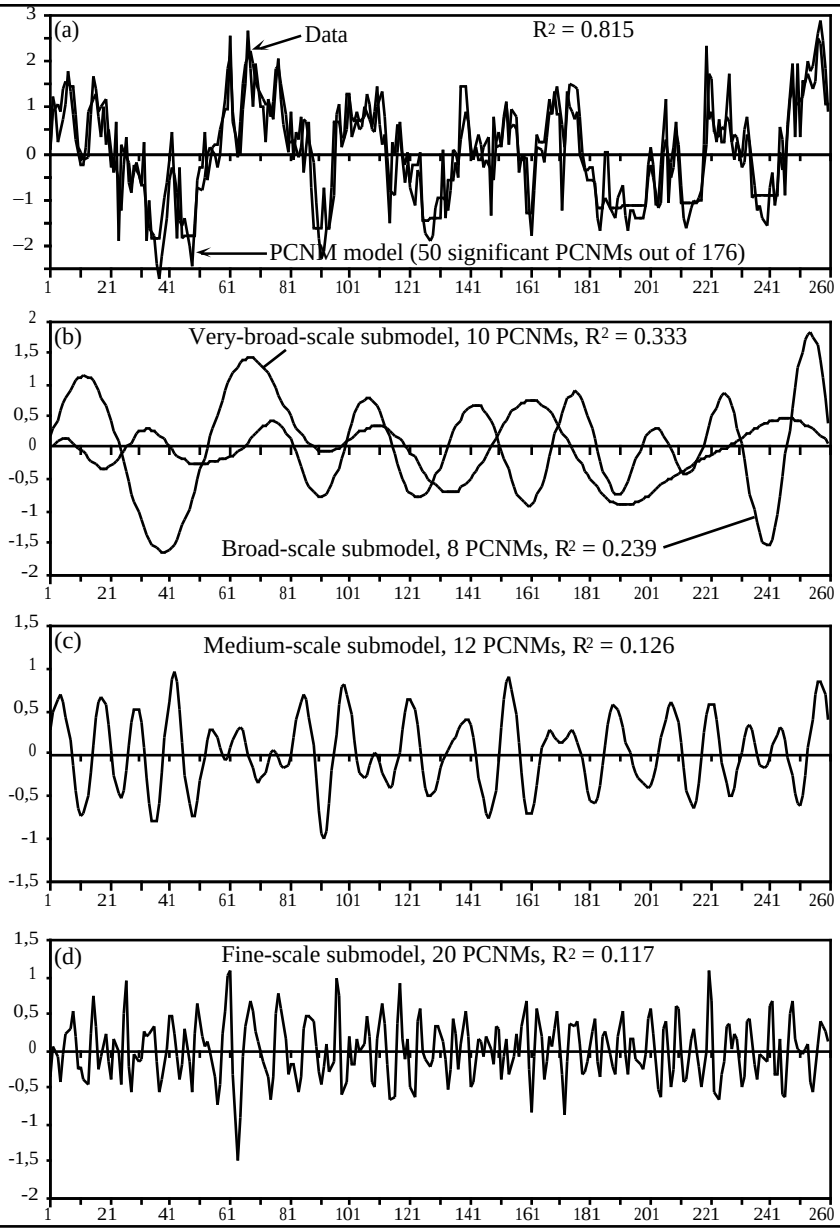
Sélection progressive

50 fonctions CPMV ont été retenues parmi 176 (test par permutations, 999 permutations).

Les CPMV ont été divisées arbitrairement en 4 sous-modèles. Les sous-modèles sont orthogonaux entre eux.

Longueurs d'onde signific. (analyse périodogramme) :

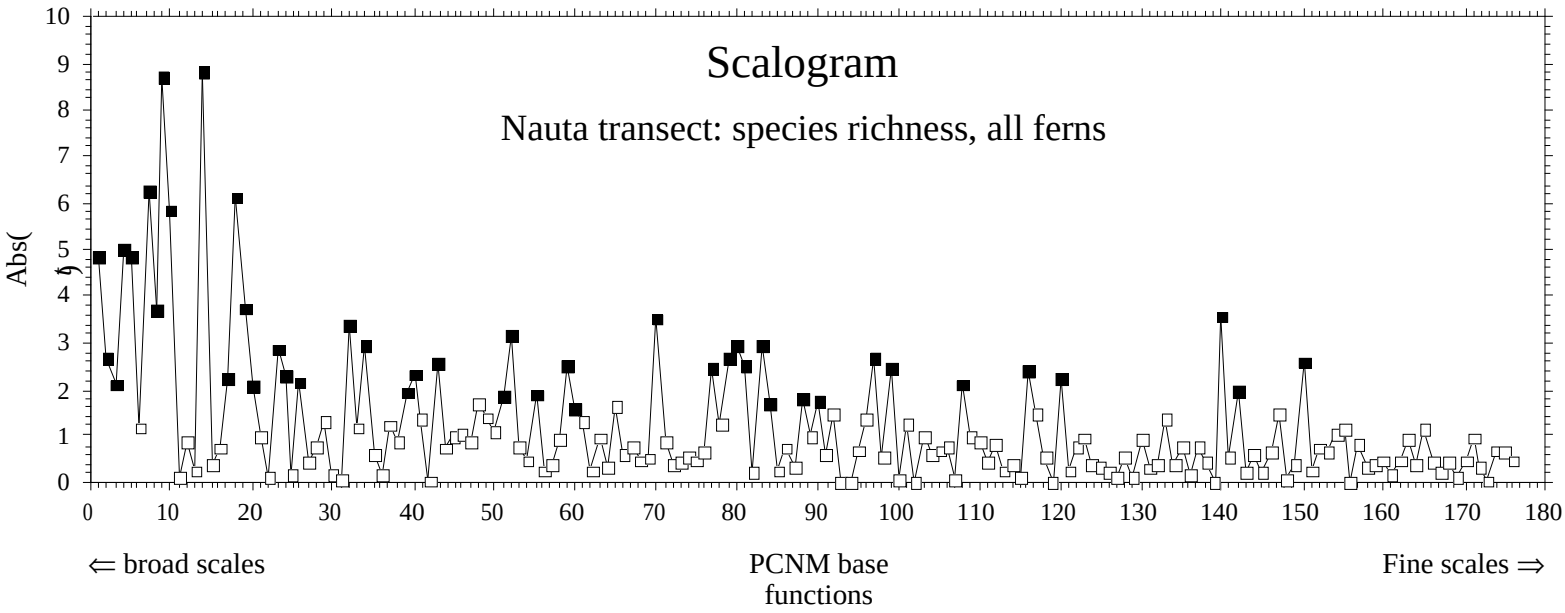
Très large : 250, 355-440 m
Large : 180 m
Intermédiaire : 90 m
Échelle fine : 50, 65 m



Interprétation : régression sur les variables environnementales

	Adiantum tomentosum	Very broad	Broad	Medium	Fine
R^2 of PCNM submodel on *A. tomentosum*		0.333	0.239	0.126	0.117
R^2 of envir. on submodel		0.347	0.334	0.157	
R^2 of envir. on *A. tomentosum*	0.436	0.116	0.080	0.020	
Elevation (m)	< 0.0001	< 0.0001	< 0.0001		
Thickness of soil organic horizon (cm)	0.0004	< 0.0001			
Waterlogging	0.0393			0.0007	
Canopy height (m)	0.0011	< 0.0001			
Canopy coverage (%)				0.0024	
Shrub coverage (%)	0.0621	0.0341		0.0001	
Herb coverage (%)		0.0002			
Trees 3 - 7.5 cm DBH	0.0083		0.0009		
Lianas 3 - 7.5 cm diameter				0.0413	
Lianas 8 - 15 cm diameter				0.0183	

Scalogramme de la richesse en espèces, toutes les espèces de fougères



Exemple 2

Échantillonnage d'une surface sur une grille régulière
Chlorophylle *a* dans un étang saumâtre[1]

Données : chlorophylle *a* à 63 sites dans un étang (surface géogr.).

Plan d'échantillonnage : 63 sites formant une grille régulière (espacement de 1 km) dans l'étang de Thau (19 km x 5 km), France.

Questions

- À quelles échelles spatiales la chlorophylle *a* est-elle structurée ?
- Ces échelles sont-elles reliées à certaines var. environnementales ?

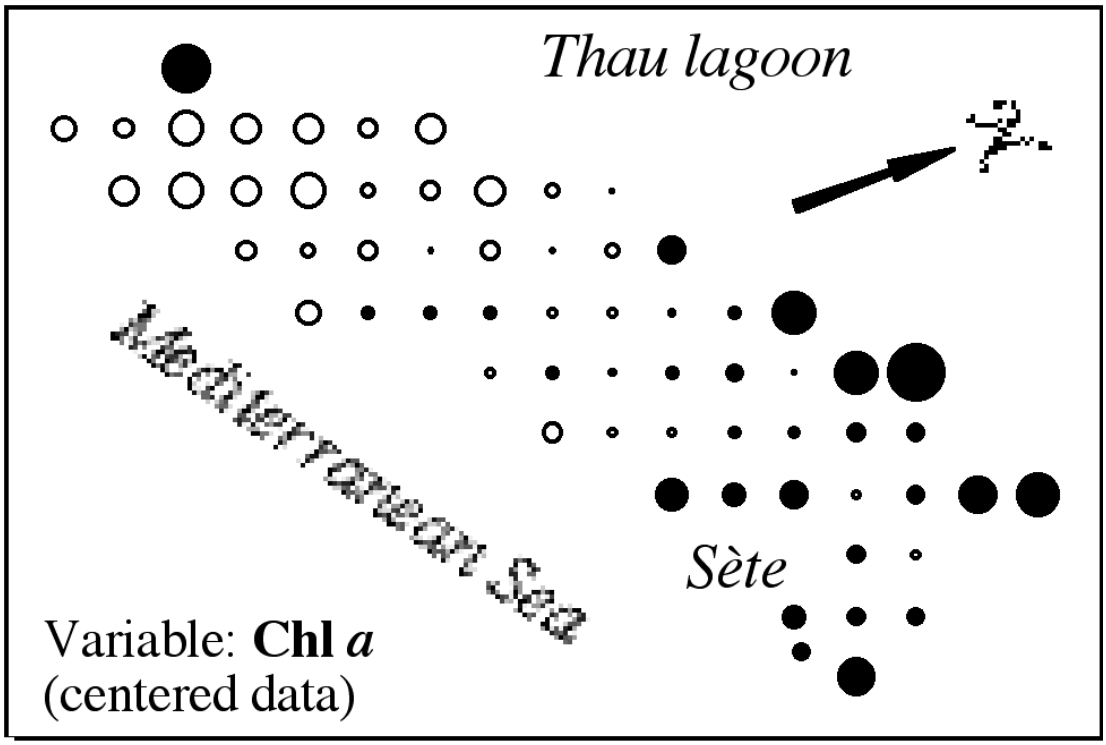
Pré-traitement

- Aucun.

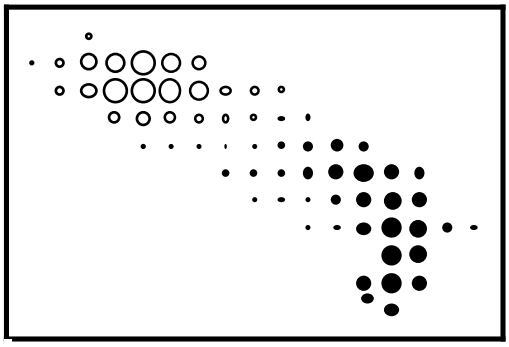
Sélection progressive

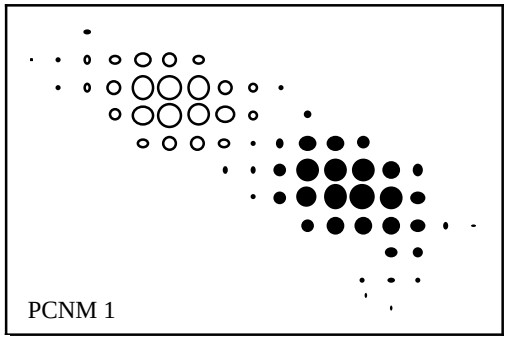
- 12 fonctions CPMV ont été retenues, parmi 45.

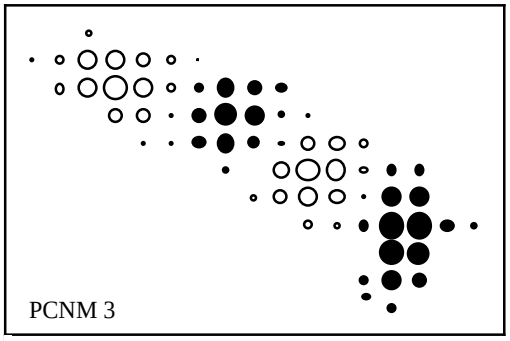
[1] Données analysées d'abord par Legendre & Troussellier 1988; ré-analysées par Borcard et al. 1992, Borcard, Legendre, Avois-Jacquet & Tuomisto 2004 et Legendre & Borcard 2005.

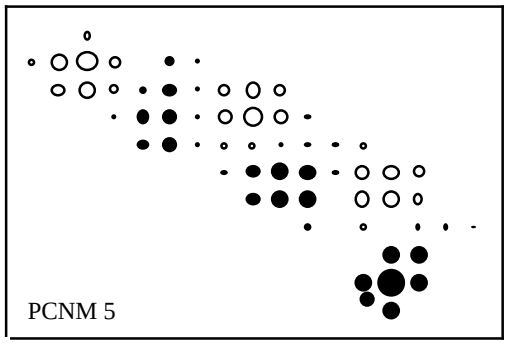

Thau lagoon
Mediterranean Sea
Sète
Variable: **Chl *a***
(centered data)

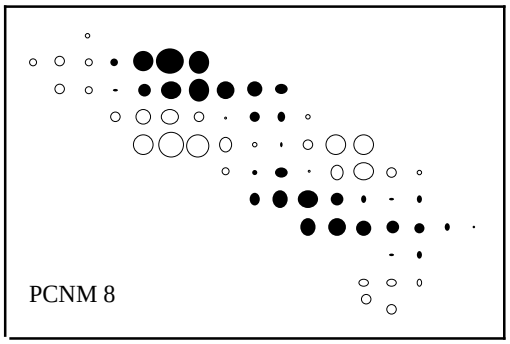
Chl *a*
Broad-scale model
(PCNM 1, 3, 5, 8)
$R^2 = 0.410$



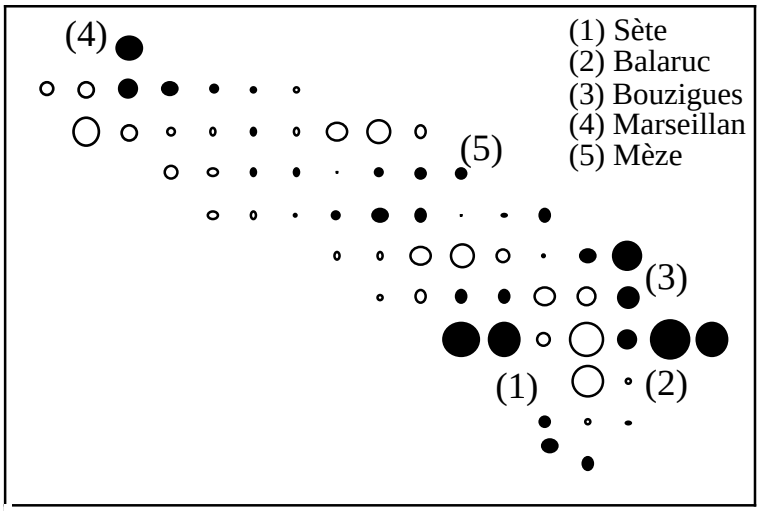




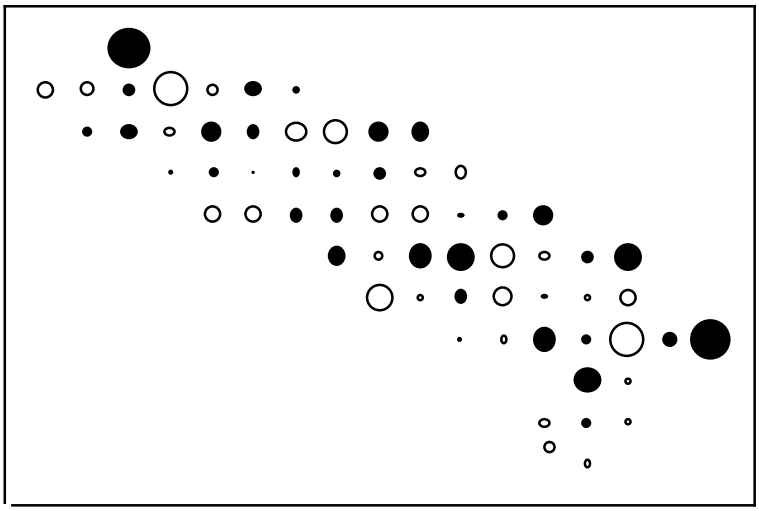




Chl *a*
Intermediate-scale model
(PCNM 13, 14, 17, 19, 20)
$R^2 = 0.235$



Chl *a*
Fine-scale model
(PCNM 24, 28, 36)
$R^2 = 0.135$



Exemple 3

Poissons (données multivariables) dans la zone littorale d'un lac[1]

Données : composition de la communauté de poissons (7 espèces).

Plan d'échantillonnage : la zone littorale du lac fut divisée en 90 sites contigus. Recensement visuel par deux plongeurs en juin 2001.

Questions

- À quelles échelles spatiales la communauté est-elle structurée ?
- Ces échelles sont-elles reliées à celles des var. environnementales ?

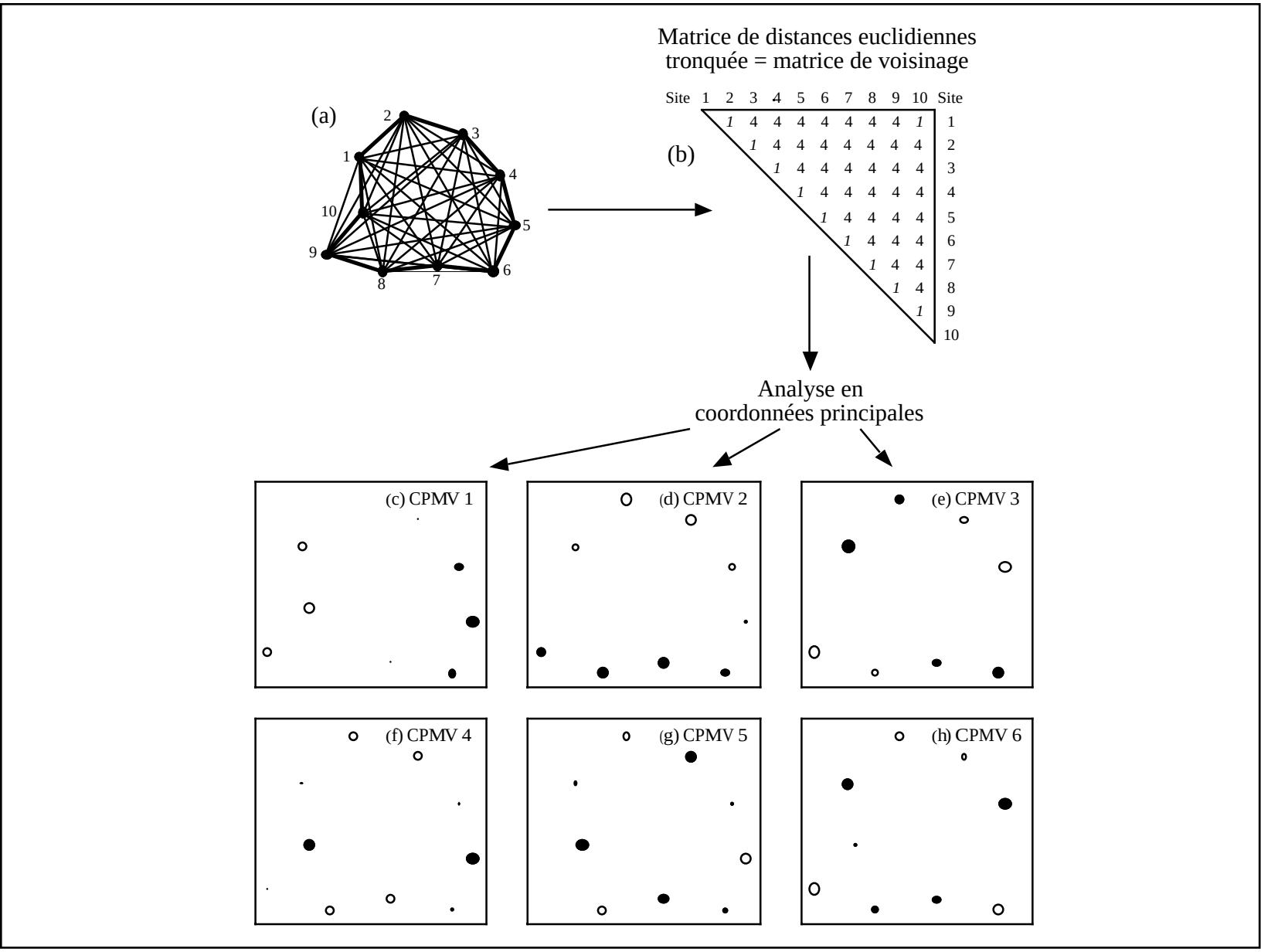
Pré-traitement

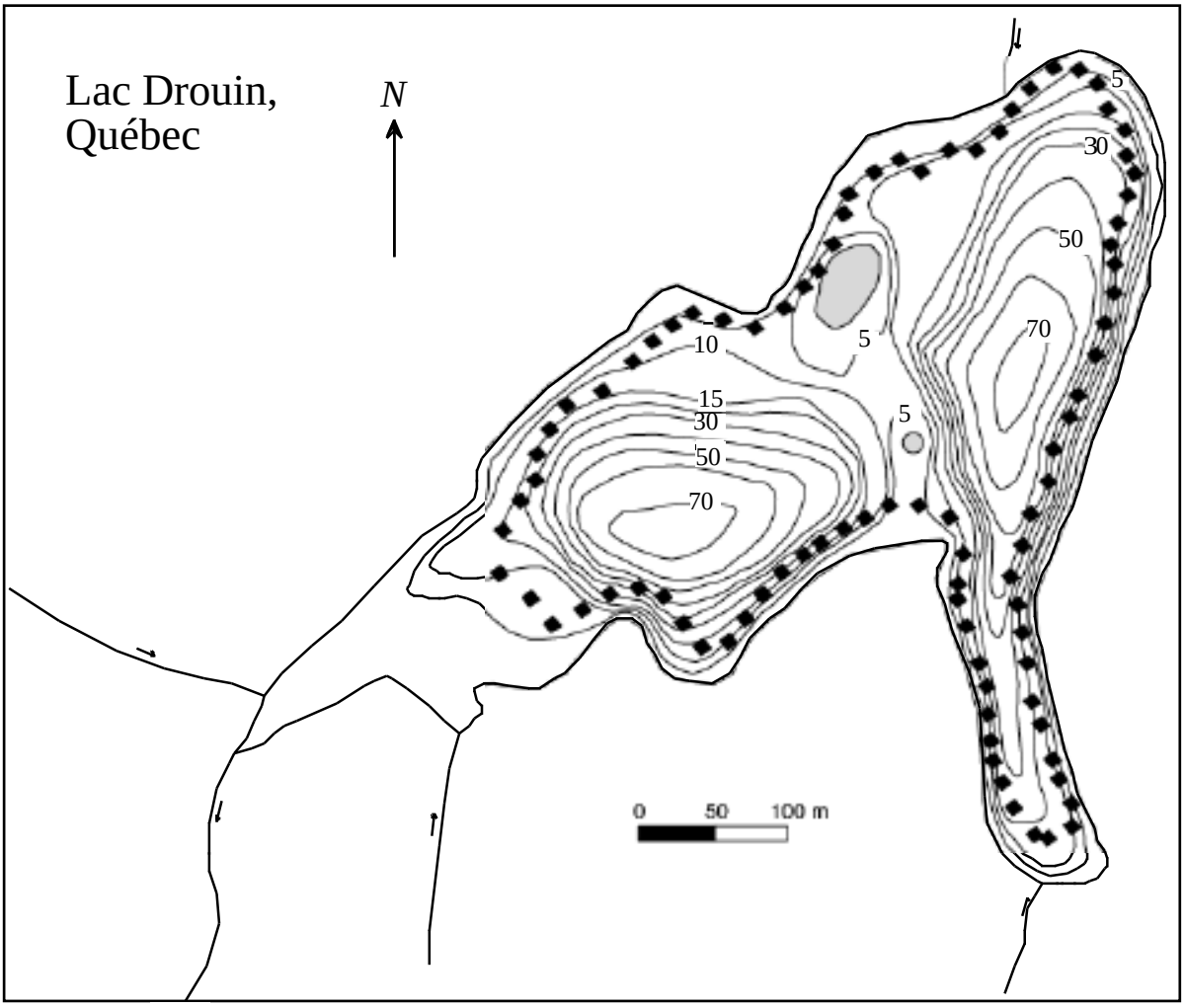
- Aucun.

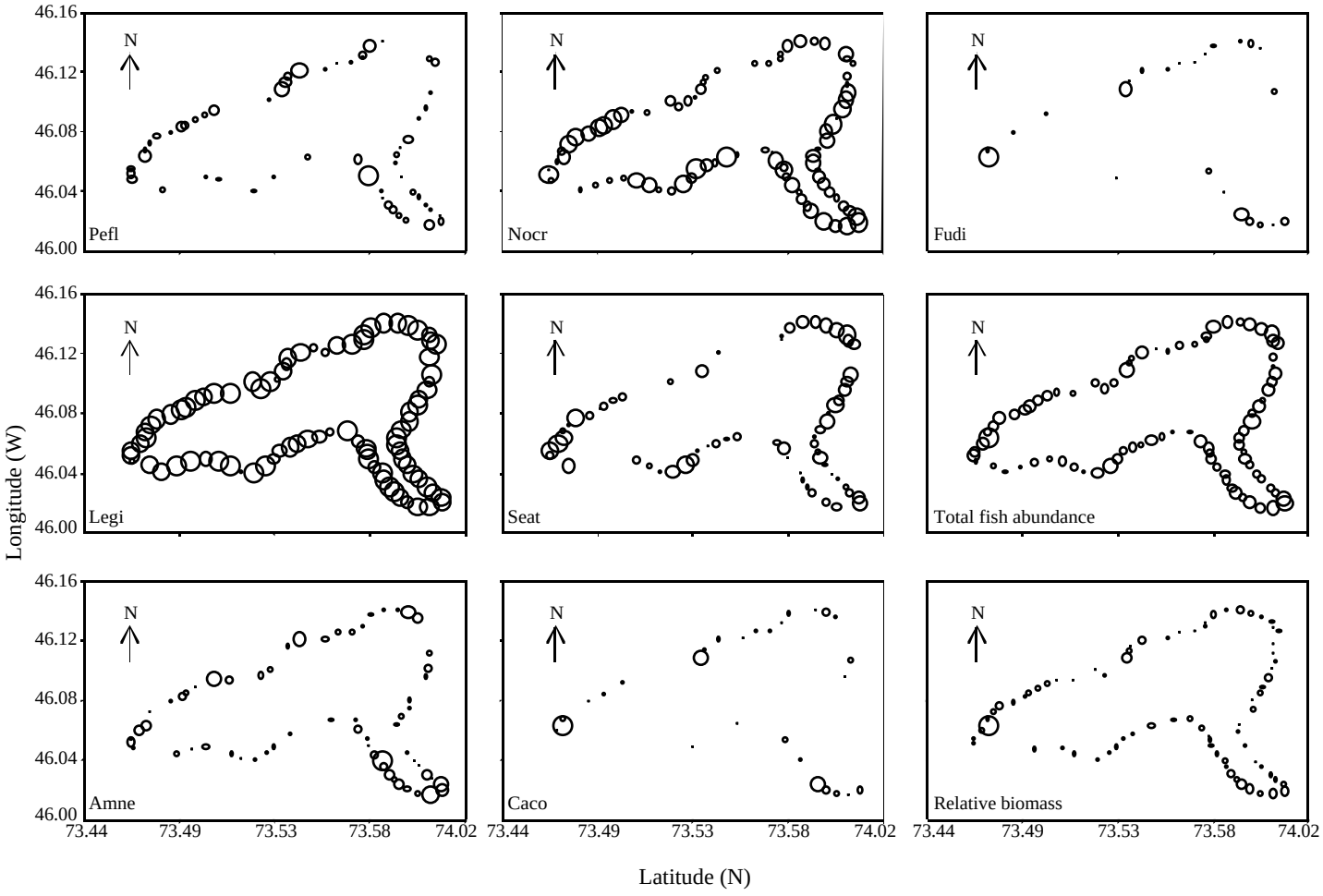
Sélection progressive

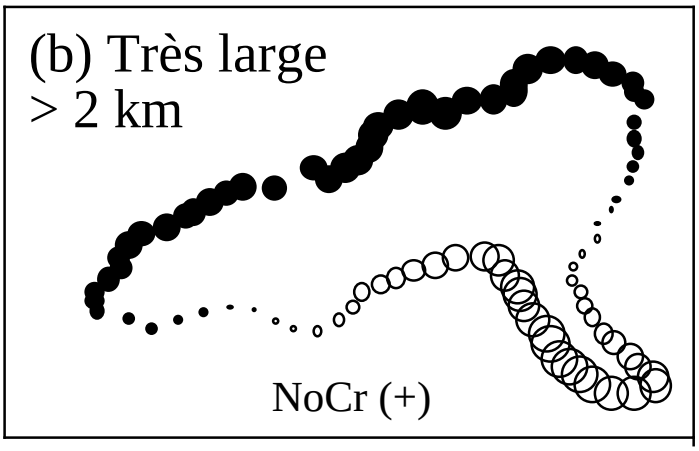
- 15 fonctions CPMV furent retenues, parmi 60.

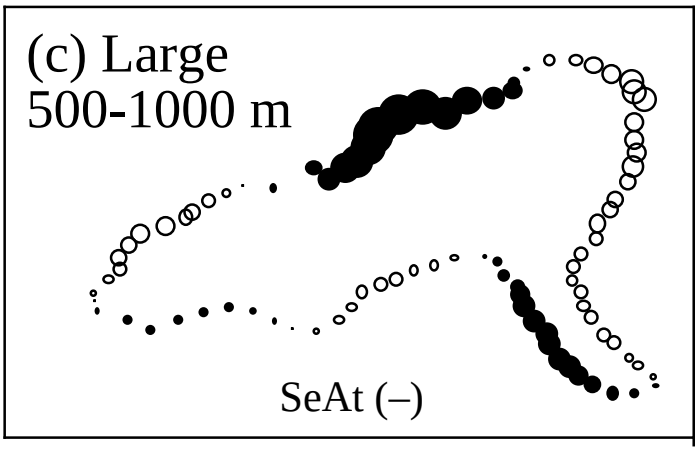
[1] Brind'Amour, A., D. Boisclair, P. Legendre & D. Borcard. 2005 Multiscale spatial distribution of a littoral fish community in relation to environmental variables. *Limnology and Oceanography* 50: 465-479.

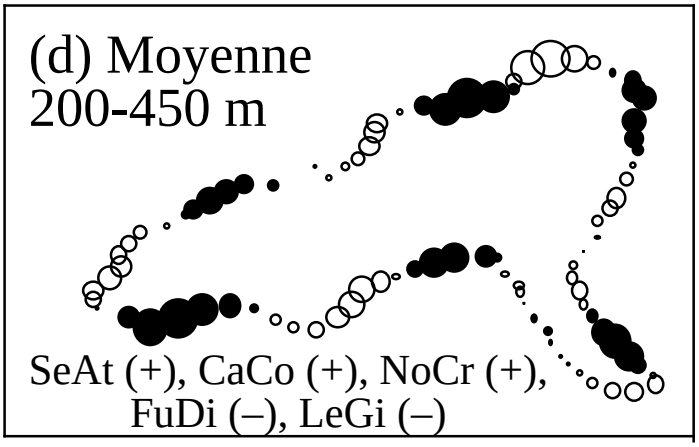

Matrice de distances euclidiennes
tronquée = matrice de voisinage
(a)
(b)
Site 1 2 3 4 5 6 7 8 9 10 Site
1 4 4 4 4 4 4 4 1 1
1 4 4 4 4 4 4 4 2
1 4 4 4 4 4 4 3
1 4 4 4 4 4 4
1 4 4 4 4 5
1 4 4 4 6
1 4 4 7
1 4 8
1 9
10
Analyse en
coordonnées principales
(c) CPMV 1
(d) CPMV 2
(e) CPMV 3
(f) CPMV 4
(g) CPMV 5
(h) CPMV 6

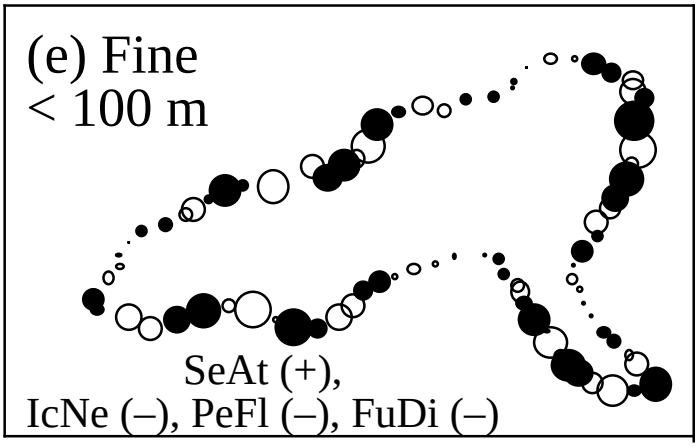

Lac Drouin,
Québec
N
5
30
50
70
10
5
15
30
50
70
5
0
50
100 m


Longitude (W)
46.16
46.12
46.08
46.04
46.00
N
Pefl
Nocr
Fudi
Legi
Seat
Total fish abundance
Amne
Caco
Relative biomass
73.44
73.49
73.53
73.58
74.02
Latitude (N)









Code des espèces : CaCo, *Catostomus commersoni* (meunier noir) ; FuDi, *Fundulus diaphanus* (fondule barré) ; IcNe, *Ictalurus nebulosus* (barbotte brune) ; LeGi, *Lepomis gibbosus* (crapet-soleil) ; NoCr, *Notemigonus crysoleucas* (chatte de l'est) ; PeFl, *Perca flavescens* (perchaude) ; SeAt, *Semotilus atromaculatus* (mulet à cornes).

Inventaire de juin 2001 au lac Drouin : coefficients de régression centrés réduits (*b*) des variables environnementales qui contribuent de façon significative à expliquer les patrons spatiaux (modèles CPMV) de la communauté de poissons à trois échelles différentes.

	Community composition			Total fish density		Relative fish biomass		
	V. broad	Broad	Meso	Broad	Meso	Broad	Meso	Fine
R^2=	0.513***	0.263***	0.198***	0.171***	0.047*	0.363*	0.295*	0.052*
Riv		–0.247**		–0.211*		–0.409***		
Lit		–0.376***						
Z								
S1		–0.214*				–0.226*		
S3								
S4				–0.220*		–0.353***		
S5								
S8	–0.167*			0.279**				
U1								
U2	0.195*						0.193*	
U3								
U4								
U5								
Tree					-0.217*			
Sub								
Emer	0.325***		–0.292**				0.195*	
Cov	0.232**		0.304**				–0.371***	
Fet	0.565***		0.262*				–0.292**	
Trib							0.315***	–0.228*

Dans ces trois exemples, nous avons procédé comme suit :

- Analyse CPMV du vecteur ou tableau-réponse **Y** ;
- Division des fonctions CPMV en sous-modèles ;
- Interprétation des sous-modèles par des variables explicatives.

Les fonctions CPMV peuvent également être utilisées dans une *partition de la variation*[1]. On pourrait, par exemple, partitionner la variation de **Y** en fonction d'un tableau explicatif environnemental **X** et de plusieurs tableaux contenant des sous-modèles de fonctions CPMV: $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{W}_3$ etc.

[1] Borcard & Legendre (2002) en donnent un exemple.

Références

Documents disponibles en PDF à l'adresse http://www.bio.umontreal.ca/legendre/

Programme qui crée les fonctions de base CPMV: Borcard, D. and P. Legendre. 2004. SpaceMaker2 – User's guide. Département de sciences biologiques, Université de Montréal. 20 pages.

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

Borcard, D. & P. Legendre. 1994. Environmental control and spatial structure in ecological communities: an example using Oribatid mites (Acari, Oribatei). *Environmental and Ecological Statistics* 1: 37-61.

Borcard, D. & P. Legendre. 2002. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling* 153: 51-68.

Borcard, D., P. Legendre, C. Avois-Jacquet & H. Tuomisto. 2004. Dissecting the spatial structure of ecological data at multiple scales. *Ecology* 85: 1826-1832.

Legendre, P. & D. Borcard. 2005. Quelles sont les échelles spatiales importantes dans un écosystème ? In: J.-J. Droesbeke, M. Lejeune et G. Saporta (éds), *Analyse statistique de données spatiales*. Éditions TECHNIP, Paris. (Sous presse).