Bringing multivariate support to multiscale codependence analysis: assessing the drivers of community structure across spatial scales

Guillaume Guénard[1]*and Pierre Legendre[1†]

3rd February 2017

1. Département de sciences biologiques
Université de Montréal
C.P. 6128, succursale Centre-Ville
Montréal, QC, Canada H3C 3J7

1. Multiscale codependence analysis (MCA) quantifies the joint spatial distribution of a pair of variables in order to provide a spatially-explicit assessment of their relationships to one another. For the sake of simplicity, the original definition of MCA only considered a single response variable (e.g. a single species). However, that definition would limit the application of MCA when many response variables are studied jointly, for example when one wants to study the effect of the environment on the spatial organisation of a multi-species community in an explicit manner.

2. In the present paper, we generalize MCA to multiple response variables. We conducted a simulation study to assess the statistical properties (i.e. type I error rate and statistical power) of multivariate MCA (mMCA) and found that it had honest type I error rate and sufficient statistical power for practical purposes, even with

*guillaume.guenard@gmail.com (corresponding author)
†pierre.legendre@umontreal.ca

modest sample sizes. We also exemplified mMCA by applying it to two ecological data sets.

3. The simulation study confirmed the adequacy of mMCA from a statistical standpoint: it has honest type I error rates and sufficient power to be useful in practice. Using mMCA, we were able to detect variation in fish community structure along the Doubs River (in France), which was associated with large spatial structures in the variation of physical and chemical variables related to water quality. Also, mMCA usefully described the spatial variation of an Oribatid mite community structure associated with a gradient of water content superimposed on various smaller-scale spatial features associated with vegetation cover in the peat blanket surrounding Lac Geai (in Québec, Canada).

4. In addition to demonstrating the soundness of mMCA in theory and practice, we further discuss the strengths and assumptions of mMCA and describe other potential scenarios where it would be helpful to biologists interested in assessing influence of environmental conditions on community structure in a spatially-explicit way.

Language: UK English

Keywords: habitat modelling, scale, scale-dependent correlation, spatial model

Running title: Multivariate multiscale codependence analysis

Includes 5546 words, 3 tables, and 8 figures

Introduction

Multi-scale codependence analysis (MCA; Guénard et al., 2010) is a statistical method to estimate the joint spatial structures of pairs of variables by quantifying to what extent they fluctuate in unison, following the same trends in space, which are described by an orthonormal set of geographic structuring variables called spatial eigenvectors (described in particular by Griffith, 2000; Borcard and Legendre, 2002; Dray et al., 2006; Griffith and Peres-Neto, 2006; Blanchet et al., 2008). Any mention to space in the present paper may equally apply to time or space-time data and processes. These structuring variables can be calculated from regularly or irregularly-spaced points. This aspect is important for applicability to ecological data sets where sampling may often not be regular along a transect or on a grid. The interest of MCA for the analysis of ecological data lies in the fact that natural processes are almost always operating at particular spatial scales and, consequently, the ecosystem features that derive from these processes are generally structured in space. Hence, the assessment of the structures emerging from spatiotemporal organisation is now widely recognised as a cornerstone paradigm to understand ecological processes (Legendre, 1993; Wiens et al., 1993; Cottenie, 2005; Wagner and Fortin, 2005). For instance, landscape ecology is concerned about how the spatial organisation of environmental features of the landscape structures the functioning of ecosystems (Forman and Godron, 1986; Forman, 1995).

MCA was initially developed as a way of incorporating spatiotemporal information about environmental conditions in modelling the distribution of a species. In its original definition, MCA was presented as a method applicable only to single response variable.

That limitation does not reflect the impossibility of calculating multivariate codependence but, rather, a choice done in that early version of the method for the sake of simplicity. It is expected, however, that MCA could be utilised in a much broader range of applications if it could handle multivariate response data. Ecosystems are often characterized by their species content for different target groups of organisms, which are multivariate data. There is therefore a need for statistical methods that allow scientists to quantify the join spatial trends of community structure (or some other similar multivariate ecosystem response) and environmental conditions.

The objective of the present study is to develop a multivariate implementation of MCA, assess its statistical properties (type I error rate and statistical power) using a Monte-Carlo simulation study, and present a few examples of applications to help readers figure out its relevance and the practical interpretation of its results. Monte-Carlo simulations were performed for a variety of sample sizes using both parametric and permutation testing whereas the examples encompassed case scenarios from river fish ecology and wetland ecology.

Methods

Computation of multivariate MCA

To quantify the joint spatial dependence of a response and an explanatory data table, MCA requires a set of variables (**U**) suitable to represent spatial patterns of variation in the data (Guénard et al., 2010). These variables have to be centred (i.e., their values have to sum to 0) and orthonormal (i.e., their cross-product to one another $\mathbf{u}_i^\mathsf{T}\mathbf{u}_j = 0$ for all $i \neq j$, and the sum of squares $\mathbf{u}_i^\mathsf{T}\mathbf{u}_i = 1$ for all i, where $^\mathsf{T}$ denotes the matrix transpose).

Univariate multiscale codependence analysis quantifies the strength of the association between a response variable ($\mathbf{y}$) and an explanatory descriptor ($\mathbf{x}$) at a spatial scale described by a spatial eigenvector ($\mathbf{u}_i$) using a codependence coefficient $C_{\mathbf{u}_i;\mathbf{y},\mathbf{x}}$, which is the product of the (Pearson) correlation coefficients between the response variable and the spatial eigenvector with that of the explanatory descriptor and the same spatial eigenvector. When both the response variable and the descriptor are centred on their means ($\bar{y} = \bar{x} = 0$), codependence is defined as follows:

$$C_{\mathbf{y},\mathbf{x};\mathbf{u}_i} = \frac{\mathbf{u}_i^\mathsf{T}\mathbf{y}}{\sqrt{\mathbf{y}^\mathsf{T}\mathbf{y}}} \frac{\mathbf{u}_i^\mathsf{T}\mathbf{x}}{\sqrt{\mathbf{x}^\mathsf{T}\mathbf{x}}}. \tag{1}$$

To test $C_{\mathbf{u}_i;\mathbf{y},\mathbf{x}}$ for statistical significance, Guénard et al. (2010) proposed to use the τ statistic, defined as the product of two Student's t statistics corresponding to the two correlations coefficients whose product is $C_{\mathbf{u}_i;\mathbf{y},\mathbf{x}}$ (Eq. 6 in Guénard et al., 2010). From Springer (1979), the probability density function of the τ statistic corresponds to the following definite integral:

$$\tau_\nu(z) = 2\int_0^\infty \frac{t_\nu(x)\, t_\nu(z/\theta)}{\theta} d\theta, \tag{2}$$

where z is the value of the product statistic, θ is the variable to be integrated in the domain $[0, \infty]$, and $t_\nu()$ is the probability density function of Student's t distribution with ν degrees of freedom. For the purpose of the present study, we will use the abbreviation MCA$^{(\mathrm{u})}$ when referring specifically to the original method applicable to univariate response data and mMCA when referring specifically to its multivariate generalisation described below, whereas MCA will refer to either of these analyses.

To implement multivariate support in MCA, we propose to replace the left portion of Eq. 1 with the square root of the multivariate determination coefficient (R^2) of the regression between a matrix of response variables $\mathbf{Y}$ and a spatial eigenvector $\mathbf{u}_i$ as follows:

$$C_{\mathbf{u}_i;\mathbf{Y},\mathbf{x}} = \sqrt{\frac{trace((\mathbf{Y} - \mathbf{u}_i\mathbf{u}_i^{\intercal}\mathbf{Y})^{\intercal}(\mathbf{Y} - \mathbf{u}_i\mathbf{u}_i^{\intercal}\mathbf{Y}))}{trace(\mathbf{Y}^{\intercal}\mathbf{Y})}}\frac{|\mathbf{u}_i^{\intercal}\mathbf{x}|}{\sqrt{\mathbf{x}^{\intercal}\mathbf{x}}}. \tag{3}$$

where $trace()$ denotes the trace of the matrix, i.e. the sum of its diagonal elements. The sign of the right portion of Eq. 3 was discarded because it only depends on that of the relationship between $\mathbf{x}$ and $\mathbf{u}_i$, which is not really informative in the multivariate context. By extension of the τ statistic used in MCA$^{(\mathrm{u})}$ we propose to test the multivariate codependence coefficient of mMCA using the product of two Fisher-Snedecor F statistics as follows:

$$\phi_{\mathbf{u}_i \in \mathbf{U}_s;\mathbf{Y},\mathbf{x}} = (n-k-1)^2 \frac{trace((\mathbf{Y} - \mathbf{u}_i\mathbf{u}_i^{\intercal}\mathbf{Y})^{\intercal}(\mathbf{Y} - \mathbf{u}_i\mathbf{u}_i^{\intercal}\mathbf{Y}))}{trace((\mathbf{Y} - \mathbf{U}_s\mathbf{U}_s^{\intercal}\mathbf{Y})^{\intercal}(\mathbf{Y} - \mathbf{U}_s\mathbf{U}_s^{\intercal}\mathbf{Y}))}\frac{|\mathbf{u}_i^{\intercal}\mathbf{x}|^2}{(\mathbf{x} - \mathbf{U}_s\mathbf{U}_s^{\intercal}\mathbf{x})^{\intercal}(\mathbf{x} - \mathbf{U}_s\mathbf{U}_s^{\intercal}\mathbf{x})}$$

where n is the sample size, k is the number of columns of $\mathbf{U}_s$, and $\mathbf{U}_s$ is a matrix containing spatial eigenvectors that have previously been tested for significance (if any), in addition to the one being tested (i.e. $\mathbf{u}_i$). The probability density function of the ϕ statistic corresponds to the following definite integral (Springer, 1979):

$$\phi_{\nu_1,\nu_2}(z) = \int_0^{\infty} \frac{F_{\nu_1,\nu_1\nu_2}(\theta)\, F_{1,\nu_2}(z/\theta)}{\theta} d\theta, \tag{4}$$

where z is the value of the product statistic, θ is the variable to be integrated in the domain $[0, \infty]$, $F_{a,b}(...)$ is the probability density function of the Fisher-Snedecor F distribution with a degrees of freedom in the numerator and b in the denominator, ν_1 is the number of degrees of freedom corresponding to the number of linearly independent columns in $\mathbf{Y}$ (and thus the rank of $\mathrm{cov}(\mathbf{Y})$), and ν_2 is the number of residual degrees of freedom associated with the sampling sites (i.e. $\nu_2 = n - k - 1$). The assumptions related to testing $\phi_{\mathbf{u}_i \in \mathbf{U}_s;\mathbf{Y},\mathbf{x}}$ are the union of those of the multivariate regression of $\mathbf{Y}$ against $\mathbf{u}_i$ with those of the linear regression of $\mathbf{x}$ against $\mathbf{u}_i$. Notably, residuals of both $\mathbf{Y}$ and $\mathbf{x}$ with respect to $\mathbf{u}_i$ (and other eigenvectors in $\mathbf{U}_s$, if any) have to be (multivariate) normally distributed and their variances should be homogeneous along the range of values in $\mathbf{u}_i$. In cases where the normality assumption (for either $\mathbf{Y}$ or $\mathbf{x}$, or both) is not met or difficult to ascertain (e.g. when sample size is too small to reliably assess the probability distribution), testing may be done using Monte-Carlo permutations. It is also noteworthy that while the τ statistic was signed and allowed one to perform both one-way or two-way inference tests, the ϕ statistic in strictly positive and tests the null hypothesis (H_0) of no codependence against multiple two-way alternative hypotheses (i.e. H_1: presence of codependence of any sign depending on particular responses $\mathbf{y}_j$ in $\mathbf{Y}$).

The five-step testing procedure originally proposed for MCA$^{(\mathrm{u})}$ equally applies to mMCA and goes as follows:

1. Compute the vector $[C_{\mathbf{U};\mathbf{Y},\mathbf{x}}]$ of the codependence coefficients.

2. Sort values of $[C_{\mathbf{U};\mathbf{Y},\mathbf{x}}]$ in descending order.

3. Select the spatial eigenvector $\mathbf{u}_{max}$, associated with the highest codependence coefficient $C_{\mathbf{u}_{max};\mathbf{Y},\mathbf{x}}$ among those that have not been tested (i.e. $\mathbf{u}_{max}$ is not a

member of $\mathbf{U}_s$ at that point).

4. Calculate $\phi_{\mathbf{u}_i \in \mathbf{U}_s;\mathbf{Y},\mathbf{x}}$ and its associated probability (P) using the theoretical distribution or by permutation.

5. Test the significance of $\mathbf{u}_{max}$ by comparing its P-value to a predetermined significance level α. If significant, incorporate $\mathbf{u}_{max}$ permanently in $\mathbf{U}_s$ and proceed again from step 3 to test another coefficient. If non-significant, stop here.

That method ensures that we highlight the best codependence coefficients, but since many eigenvectors are generally tested (sometimes as many as the sample size minus one), it comes at the price of inflated type I error. As for MCA$^{(\mathrm{u})}$, that issue can be addressed by considering all possible inference tests as a family of independent tests (eigenvectors being orthogonal) and apply a correction to transform the probabilities of single tests (i.e. testwise P-values) in probabilities for the whole family of tests (i.e. familywise P-values). We propose using a sequential version of the Šidák correction (Šidák, 1967; Wright, 1992), the same method used by Guénard et al. (2010) for MCA$^{(\mathrm{u})}$.

Assessing goodness of fit in mMCA proceeds similarly as for MCA$^{(\mathrm{u})}$: a matrix of coregression coefficients ($\mathbf{B}_{\mathbf{U};\mathbf{Y},\mathbf{x}}$) is obtained for each response variables $\mathbf{y}_i$ (column of $\mathbf{Y}$) as follows:

$$\mathbf{B}_{\mathbf{U};\mathbf{Y},\mathbf{x}} = \left[b_{\mathbf{u}_i;\mathbf{y}_j,\mathbf{x}}\right] = \left[\frac{\mathbf{u}_i^\intercal \mathbf{y}_j}{\mathbf{u}_i^\intercal \mathbf{x}}\right],\ i = 1, 2, 3, ..., n;\ j = 1, 2, 3, ..., m, \tag{5}$$

where n is the number of spatial eigenvectors (columns of $\mathbf{U}$) and m is the number of response variables (columns of $\mathbf{Y}$; $\mathbf{B}_{\mathbf{U};\mathbf{Y},\mathbf{x}}$ has dimensions $n \times m$). Standardized coregression coefficients are similarly defined as:

$$\beta_{\mathbf{U};\mathbf{Y},\mathbf{x}} = \left[\beta_{\mathbf{u}_i;\mathbf{y}_j,\mathbf{x}}\right] = \left[\sqrt{\frac{\mathbf{x}^\mathsf{T}\mathbf{x}}{\mathbf{y}_j^\mathsf{T}\mathbf{y}_j}}\frac{\mathbf{u}_i^\mathsf{T}\mathbf{y}_j}{\mathbf{u}_i^\mathsf{T}\mathbf{x}}\right], \; i = 1, 2, 3, ..., n; \; j = 1, 2, 3, ..., m. \tag{6}$$

The function to make predictions for a new descriptor vector $\mathbf{x}_{new}$ (centred to 0 mean) is obtained by rearranging Eq. 5 as follows:

$$\mathbf{Y}_{pred}(\mathbf{x}_{new}) = \sum_{\forall i}^{\in s} \mathbf{u}_i \left\{(\mathbf{u}_i^\mathsf{T}\mathbf{x}_{new})\mathbf{b}_{\mathbf{u}_i;\mathbf{Y},\mathbf{x}}\right\}, \tag{7}$$

where s is the set of indices of the spatial eigenvectors found to be suitable to make predictions (notation $\sum_{\forall i}^{\in s}$ means "the sum for all i within set s"), while fitted values ($\hat{\mathbf{Y}}$) are obtained as an orthogonal projection of the observation $\mathbf{Y}$ unto the k-dimensional space spanned by the k selected structuring variables in set s:

$$\hat{\mathbf{Y}} = \sum_{\forall i}^{\in s} \left\{\mathbf{u}_j\mathbf{u}_j^\mathsf{T}\right\}\mathbf{Y}. \tag{8}$$

When set s is empty (i.e. no eigenvector was suitable), $\hat{\mathbf{Y}} = \mathbf{0}$ and all predicted or fitted responses are equal to their means. As in MCA$^{(\mathrm{u})}$, it is possible to use multiple descriptor variables in mMCA as long as they are involved with a mutually exclusive set of spatial eigenvectors (e.g. a descriptor $\mathbf{x}_1$ may influence $\mathbf{Y}$ following the spatial variation patterns described by $\mathbf{u}_1$ and $\mathbf{u}_3$ at the same time as a descriptor $\mathbf{x}_2$ influences $\mathbf{Y}$ following spatial patterns described by $\mathbf{u}_2$ and $\mathbf{u}_4$, but $\mathbf{x}_2$ cannot be involved with either $\mathbf{u}_1$ or $\mathbf{u}_3$ because $\mathbf{x}_1$ has already taken them). That exclusiveness condition guarantees that the component of the response brought by the different descriptors are orthogonal and can be combined in an additive manner.

Simulation study

We ran Monte-Carlo simulations to estimate the type I and II error rates (i.e., the probability of rejecting the null hypothesis when it is true and that of failing to reject it when it is false, respectively) generated by mMCA when it was applied to pairs of variables $\mathbf{Y}$ (multivariate) and $\mathbf{x}$ (univariate). Simulations were performed using parametric testing for normal random deviates and by permutation testing for non-normal random deviates simulating species abundances. These non-normal deviates were generated as the floor-rounded integers of the exponential of random normal deviates with mean of 0 and standard deviation 1.5. That approach generated a zero-inflated distribution. We regarded that distribution as a fair approximation of that often encountered for species abundances in the wild.

The procedure consisted in generating transects of N evenly spaced sampling locations, by assigning sets of pseudo-random numbers to an $N \times M$ response data matrix $\mathbf{Y}$ and to a descriptor vector $\mathbf{x}$ with N elements. We used seven different sample sizes (N) between 10 and 1 000, which we each combined with four different numbers of species (M) between 1 and 500 (Table 1), resulting in 28 different conditions which were all analysed using parametric tests, whereas samples with sizes up to 100 were also analysed using permutations tests. The grand total of simulated conditions, including those with parametric and permutation tests, was therefore 44. Each conditions was tried 10 000 times; 440 000 simulations were thus done.

Each simulation trial consisted in testing the pseudo-random data set for the statistical significance of a single, randomly-picked spatial eigenvector. The resulting

P-value was used to assess type I error rate. Then, we took the fitted values associated with the spatial eigenvector tested previously ($\hat{\mathbf{Y}}$ and $\hat{\mathbf{x}}$), standardized them to a variance of 1, added to them some amount of normally-distributed pseudo-random deviates with mean 0 and variance 1 ($\mathfrak{N}(0, 1)$), and tested the resulting variables (referred to as $\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{x}}$, respectively) to assess, this time, the type II error rate. The amount of noise added to the fitted values was chosen by independently drawing two pseudo-random numbers between 0 and 1. The first number was used to set the signal-to-noise ratio (snr) of the trial as $snr = \frac{r_1}{\sqrt{1-r_1^2}}$, while the second number was used to distribute the snr between $\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{x}}$ as follows:

$$\tilde{\mathbf{Y}}_c = r_1 r_2 \, \hat{\mathbf{Y}} + \sqrt{1-r_1^2}\sqrt{1-r_2^2}\, \mathfrak{N}(0,1), \tag{9}$$

$$\tilde{\mathbf{x}}_c = r_1 \sqrt{1-r_2^2}\, \hat{\mathbf{x}} + \sqrt{1-r_1^2 r_2^2}\, \mathfrak{N}(0,1). \tag{10}$$

That approach is similar to that used by Guénard et al. (2010) to assess the type II error rate of MCA$^{(\mathrm{u})}$, with adaptations to multiple response variables. Here, we standardized the total variance of the fitted response matrix ($\hat{\mathbf{Y}}$) to a value of 1 prior to their combination with random deviates, but let the variances fluctuate among individual columns ($\hat{\mathbf{y}}_j$). Also, to obtain numerically exact snr values, Guénard et al. (2010) used random deviates with variance of exactly 1. We deemed that last step unnecessary in the present study since the snr value is never known in real data; we can only obtain an estimate of the population variance rather than its exact value.

Illustrative examples

We used two well-studied data set to illustrate the application of mMCA. The first data set was collected by Verneaux (1973) and consists of 30 sites sampled along a 453 km transect in the Doubs, a river located in eastern France, in which 27 fish species were observed (the response variables) and 11 explanatory quantitative variables (the descriptors) were measured. These descriptors were the river slope (*slope*, ‰), mean minimum discharge ($flow$, $\mathrm{m^3s^{-1}}$), pH, hardness ($hardness$, i.e. Calcium concentration, $\mathrm{mg\,L^{-1}}$), biological oxygen demand (BOD, $\mathrm{mg\,L^{-1}}$), dissolved phosphate ($\left[\mathrm{PO_4^{3-}}\right]$, $\mathrm{mg\,L^{-1}}$), nitrate ($\left[\mathrm{NO_3^-}\right]$, $\mathrm{mg\,L^{-1}}$), ammonium ($\left[\mathrm{NH_4^+}\right]$, $\mathrm{mg\,L^{-1}}$), and oxygen ($[\mathrm{O_2}]$, $\mathrm{mg\,L^{-1}}$). Spatial eigenfunctions were calculated on the basis of the distance from the source of the river (in km, as the fish swim). Fish count data were Hellinger-transformed into square-rooted profiles of relative species abundances before analysis. As in previous published studies of this data set, site 8, where no fish were caught, was excluded from our analysis.

The second data set was collected by Borcard and Legendre (1994) and consisted of 70 cores mostly consisting of *Sphagnum* mosses, sampled from a rectangular plot, approximately $2.5\,\mathrm{m} \times 10\,\mathrm{m}$, located on the peat mat surrounding Lac Geai, which is a bog lake located on the territory of the Station de Biologie de l'Université de Montréal in Saint-Hippolyte, Québec, Canada (lat. +45.9954, lon. −73.9936). The data set consists of a response table of abundances (counts) of 35 morpho-species of Oribatid mites (Acari) and a second table containing five environmental variables: two quantitative (substratum density, $\mathrm{g\,L^{-1}}$; water content of the substratum, in % of volume) and three qualitative (substrate composition, seven classes; presence and abundance of shrubs, three ordered classes; micro-topography the peat, two classes). For the analysis, the qualitative variables were transformed into 12 (i.e. $7+3+2$) binary (dummy) variables, yielding a grand total of 14 descriptors. Spatial eigenfunctions were calculated using the geographic (i.e. Euclidean) distances between the sampling sites in the rectangular plot.

Computer package

The R computer package "codep" that was originally developed for MCA$^{(u)}$ as been updated to support mMCA from version $0.6-5$ onward. It is freely available online for multiple computer platforms from the Comprehensive R Archive Network (CRAN: `https://cran.r-project.org/`).

Results

Simulation study

Type I error rate

The type I error rates obtained from the simulation study were close to the significance levels of the test. The expected rejection values under the null hypothesis of absence of codependence are the significance levels. This was true for all significance levels tested and all simulated sample sizes (N and M), for both the parametric (Fig. 1) and permutation (Fig. 2) tests. For $N = 10$ sites and $M = 1$ to 5 species, the permutation test was somewhat conservative, the simulations producing fewer spurious signal detection events than expected for the smallest α significance levels (0.01 and 0.005). The $N = 10$ sample size is lower than what would be found in real studies, and statistical power is extremely low under such conditions.

Statistical power

Statistical power increased as N increased (parametric test: Fig. 3; permutation test: Fig. 4), with a comparatively smaller but noticeable positive influence of M. Also, permutation tests carried out on non-normal deviates were slightly less powerful than the parametric test computed on normally-distributed data, but the method remained entirely fit for practical purposes. For instance, for a statistical power of 0.95, the permutation test detected a signal with $snr = 0.53$ for $N = 50$ and $M = 20$, and $snr = 0.96$ (a roughly equal amount of signal and random noise) for $N = 25$ and $M = 5$. For the same statistical power and under the same two (N, M) combinations, the parametric test could detect comparatively weaker signals (i.e., smaller snr) on average: 0.36 and 0.60, respectively. For species abundance data, which seldom (if ever) conform to the normal distributions, the permutation test will be the preferred method because it carries fewer assumptions than the parametric test.

Illustrative examples

Doubs River

The first sampling site was located 300 m from the source of the Doubs River and the last one was 453 km from it, with distance between neighbouring sites ranging from 1.9 to 34.4 km (average: 16.17 km). The first explanatory variable found to be significant by the mMCA test was $flow$ and it was associated with the scale of the first spatial eigenvector (that with the largest eigenvalue. The second one was BOD and it was related to the fourth spatial eigenvector. Then, $\left[NH_4^+\right]$, related to the third spatial eigenvector, and finally, $[O_2]$, at the scale of the second spatial eigenvector (Table 2).

The first principal component of the fish community structure (PC1) was positively associated with a species having preference for small and well-oxygenated streams or rivers (TRU, a Salmonid), which was found in the upstream portion of the watershed, as opposed to the more tolerant species found in large and more oxygen-depleted reaches located in the downstream portion of the watershed (Fig 5A). The sum of the four components of the spatial codependence corresponds to a slight increase in PC1 loadings in the first 150 km from the river source, followed by a steep decrease from 150 to 300 km, and a plateau from 300 km to the river mouth (Fig 5B). That figure, which shows a way of representing the influence of the MEM eigenfunctions along a river, could also be used to represent the results of mMCA analysis of transects or time series.

The second principal component (PC2), was associated with species having good tolerance to oxygen deprivation, yet showing low propensity to high $\left[NH_4^+\right]$. This was not the case for the salmonid species (TRU), which had a high positive loading on PC1. The sum of the components corresponds to a decrease in PC2 loading between 0–100 km followed by a rather sharp increase between 100–200 km, then an even sharper decrease between 200-310 km, and, finally, an increase from 310 km to the river mouth (Fig 5C).

Any other principal component associated to a substantial portion of the community variation could have been analysed similarly with respect to spatial codependence.

Oribatid mites

The strongest component of multiscale codependence associated peat water content ($WaterCont$) with the Oribatid community structure at the scale of the first spatial eigenvector (MEM1; Table 3). The latter covers the whole study plot in the north-south direction (i.e., from the forest in the south to the northern edge where the peat mat meets the open lake water). The second strongest component associated community structure with the prevalence of shrubs ($Shrub : Many$) at the spatial scale described by the fourth spatial eigenvector (MEM4), which also varies in the north-south direction along the plot, forming a pair of waves having roughly half the wavelength of MEM1. The third component associated community structure with the first type of peat moss assemblage ($Subs : Sphagnum1$; peat containing *Sphagnum rubellum* with some *S. magellacinum*) at the scale of the second spatial eigenvector (MEM2), which describes a wave with similar wavelength and orientation as MEM1, but offset by approximately a quarter of a wavelength ($\approx 90°$). The fourth and last statistically significant component of multiscale codependence pinpoints hummock ($Topo : Hummock$, i.e. elevated landforms) as another driver of Oribatid community structure at the scale of the third spatial eigenvector (MEM3). MEM3 varies transversely with respect to the north-south geographic axis of the plot.

Morpho-species with positive loadings on the first principal component of the mite community structure (PC1; e.g., Sp16, Sp31; Fig. 6) are found in peat with high water content, few shrubs, while having association with substrate composed with *Sphagnum rubellum* with some *S. magellacinum* and elevated peat mounds (Fig. 7). They oppose to the species with negative PC1 loadings (e.g., morpho-species Sp13, Sp14, Sp15). The combination of all these separate effects highlight that a large amount of species variation occurs along an edaphic gradient associated with wetter substrate as one approaches the open lake water.

On the other hand, species with positive loadings on the second principal component of the mite community structure (PC2; e.g., morpho-species Sp13, Sp16, Sp23; Fig. 6) are found in smaller abundances in peat with high water content, but follow similar trends with respect to the other descriptors, preferring few shrubs, $Subs : Sphagnum1,$ and elevated peat mounds, compared to morpho-species with negative PC2 loadings (e.g., Sp31, Fig. 8). The combination of these effects highlights the fact that species were distributed along an axis partially inclined east-west with respect to PC1. This is likely to be due to the fact that species with high positive PC2 are more prevalent in sites on peat mounds, which are more prevalent east of the plot, compared to those with high positive PC1 loadings.

Discussion

In the present study, we defined an extension of multiscale codependence analysis for multivariate response data sets, and investigated its statistical properties. The method performed as expected, yielding honest inference tests (i.e. correct type I error) and having good statistical power, even for relatively modest sample sizes compared to those generally encountered in community ecology. Adding species improved statistical power, but not as much as adding sampling sites. In that respect, our simulation study was sufficiently extensive, covering a wide range of conditions, to provide a clear demonstration that multivariate MCA (mMCA) is a useful method for practical statistical analysis.

The three main assumptions underlying mMCA with parametric tests include 1) multinormality of the residuals of the response against the spatial eigenvectors involved as well and normality of the residuals of the explanatory variables against these eigenvectors, 2) linear relationships between the response and the eigenvectors and between the descriptors and the eigenvectors, and 3) homogeneity of the residuals' variances (i.e. homescedasticity). Permutation testing relaxes the normality assumptions, leaving assumptions 2 and 3 to be satisfied. In the present study, we did not assess the robustness of the method when these assumptions are not met. Another future development to mMCA would consist in generalizing the method for other frequency distributions in the exponential family using Iteratively Re-weighted Least Squares (IRLS), as in Generalised Linear Models (GLM; Nelder and Wedderburn, 1972; Hastie and Pregibon, 1991). Calculations would proceed as in the normally-distributed case described in the Methods section, but with IRLS weights.

Fish assemblages in the Doubs were driven by flow quality, which varied following the river's course main gradient, but also by chemical conditions related to water quality (namely BOD, $\left[\mathrm{NH}_4^+\right]$, and $[\mathrm{O}_2]$), which varied following large-scale successions. The Brown trout (TRU) was the species most responsive to these effect. The analysis highlighted that this species was positively associated to NH_4^+-rich waters in spite of its well-known reliance on high concentrations of dissolved oxygen. NH_4^+ is the form of nitrogen that is readily produced by fish through excretion. However, under aerobic conditions any NH_4^+ is rapidly oxidized to ammonia (NH_3), nitrite (NO_2^-), and finally NO_3^- by ubiquitous bacteria. NO_2^-, which is the intermediate in the nitrification process, is toxic to fish as it binds to haemoglobin and hinders oxygen transport (see Lewis and Morris, 1986 for a review) and salmonids are among the most sensitive fish to that anion. Local conditions affecting the nitrification process by slowing the conversion of NH_4^+ to toxic NO_2^- may explain that association between $\left[NH_4^+\right]$ and fish community structure. A larger study involving more extensive sampling may help shed light on the effect of nitrification on fish assemblages in river ecosystems.

Oribatid mite assemblages in the peat mat surrounding Lac Geai were primarily driven by the peat's water content, which varied widely following a gradient going from the open water (north) towards the forest edge (south), and then by the presence of dense shrubs. The effects of peat moss assemblages and landforms were also evidenced. The mite morpho-species responded in various ways to variation in their habitat structure, probably as a consequence of their traits, such as their ability to move up and down in the peat mat, their preferred sources of food, and multiple physiological requirements. Had we had information about traits for the different species in that data set, it would have been computationally straightforward to project them on the principal components for the sake of displaying their prevalence in different parts of the sampling plot. In that respect, a future development of the codependence method may involve quantifying the spatially-explicit relationships between species traits and environmental variables (e.g., using bilinear algebra) instead of the relationships between multiple species responses and the environment, as illustrated in the present study.

It is noteworthy that it is possible to nest many local-scale mMCAs within a single analysis performed at a larger spatial scale. For instance, one may want to analyse the local and regional patterns of codependence for a mosaic of forested patches spread at the regional scale in a landscape. Assuming that each forested patch was sampled at multiple locations, one could perform mMCAs on each patch and then nest these local mMCAs in a single, regional, mMCA. However, if only a few locally-repeated measurements are available to perform local mMCAs with reasonable statistical power (e.g. $N < 20$ for the local samples), one should perform a single mMCA.

In the later patchwork scenario, within-patch distances are much smaller than among-patch distances. As a consequence, there is a gap between the smallest patterns of regional spatial variation and the largest patterns of local spatial variation. When a single mMCA is used, representing scales of either the regional or local spatial variation in a discrete fashion, using a set of spatial eigenvectors specially tailored for that purpose, gives results that are easier interpret compared to using a single set of spatial eigenvectors. It can be achieved by first calculating regional-scale spatial eigenvectors, substituting the patch centroid for individual observations. That analysis yields a maximum of $N_p - 1$ non-zero eigenvalues (where N_p corresponds to the number of forested patches), their associated eigenvectors being invariant among the sites pertaining to a patch. Then, one can calculate the local spatial eigenvectors for each patch. Each of these sets has to be padded to match the size of the whole data set, by assigning the value 0 to the elements corresponding to the observations in the other patches, as shown in Appendix 1 of Declerck et al. (2011). The local eigenvector sets thus padded are appended to the regional eigenvectors. One computes the cumulative sum of the eigenvalues in the same order as the eigenfunctions are appended. From that procedure, the maximum number of local eigenvectors one can obtain is $N - N_p$, where N corresponds to the total number of sites in all the patches. That number adds to that of the regional eigenvectors to give a great maximum of $N - 1$ spatial eigenvectors. That number is the same as the maximum number of eigenvectors not accounting for the spatial scale gap associated with the spatial organization of the patches in the landscape. Other examples where such spatial arrangement can be observed are lakes in a landscape, islands of an archipelago, coral reefs, etc.

We are hoping to see many application of mMCA in the near future given its usefulness to ecologists and environment scientists interested in unveiling the role of the naturally-occurring and anthropogenic phenomena structuring the spatial distribution of species assemblages and other environmental responses in the landscape. The now impressive number of large-scale (and often geographically referenced) data set now being publicly available on the Internet is an opportunity to revisit many hypotheses that might have been left untested by previous studies. The method allows researchers to readily test hypotheses that could not have been directly tested before, which may allow previously overlooked theories about the functioning of nature to emerge.

Acknowledgments

We are thankful to the many people who helped us during the elaboration of the present study by providing useful comments and suggestions. G. Guénard was supported by Discovery Grant #7738 from the Natural Sciences and Engineering Research Council of Canada (NSERC) to P. Legendre.

References

Blanchet, F. G., Legendre, P., and Borcard, D. (2008). Modelling directional spatial processes in ecological data. *Ecol. Model.*, 215:325–336.

Borcard, D. and Legendre, P. (1994). Environmental control and spatial structure in ecological communities: an example using Oribatid mites (Acari, Oribatei). *Environ. Ecol. Stat.*, 1:37–61.

Borcard, D. and Legendre, P. (2002). All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecol. Model.*, 153:51–68.

Cottenie, K. (2005). Integrating environmental and spatial processes in ecological community dynamics. *Ecology Letters*, 8:1175–1182.

Declerck, S. A. J., Coronel, J. S., Legendre, P., and Brendonck, L. (2011). Scale dependency of processes structuring metacommunities of cladocerans in temporary pools of High-Andes wetlands. *Ecography*, 34:296–305.

Dray, S., Legendre, P., and Peres-Neto, P. (2006). Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecol. Modelling*, 196:483–493.

Forman, R. T. T. (1995). *Land Mosaics: The Ecology of Landscapes and Regions.* Cambridge University Press, Cambridge, UK.

Forman, R. T. T. and Godron, M. (1986). *Landscape Ecology.* John Wiley and Sons, Inc., New York, NY, USA.

Griffith, D. A. (2000). A linear regression solution to the spatial autocorrelation problem. *J. Geograph. Syst.*, 2:141–156.

Griffith, D. A. and Peres-Neto, P. R. (2006). Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses. *Ecology*, 87:2603–2613.

Guénard, G., Legendre, P., Boisclair, D., and Bilodeau, M. (2010). Multiscale codependence analysis: an integrated approach to analyze relationships across scales. *Ecology*, 91:2952–2964.

Hastie, T. J. and Pregibon, D. (1991). *Generalized linear models*, volume Statistical models in S, chapter 6, pages 195–247. Wadsworth, Pacific Grove, CA.

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

Lewis, W. M. and Morris, D. P. (1986). Toxicity of nitrite to fish: A review. *Trans. Am. Fish. Soc.*, 115:183–195.

Nelder, J. and Wedderburn, R. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A*, 135:370–384.

Springer, M. D. (1979). *The algebra of random variables.* John Wiley & Sons Inc., Hoboken, NJ, USA.

Verneaux, J. (1973). *Cours d'eau de Franche-Comté (Massif du Jura). Recherches écologiques sur le réseau hydrographique du Doubs. Essai de biotypologie.* Thèse d'état, Besançon, France.

Šidák, Z. (1967). Rectangular confidence regions for means of multivariate normal distributions. *J. Am. Stat. Ass.*, 62:626–633.

Wagner, E. H. and Fortin, M. J. (2005). Spatial analysis of landscapes: concepts and statistics. *Ecology*, 86:1975–1987.

Wiens, J. A., Stenseth, N. C., Van Horne, B., and Ims, R. A. (1993). Ecological mechanisms and landscape ecology. *Oikos*, 66:369–380.

Wright, P. S. (1992). Adjusted p-values for simultaneous inference. *Biometrics*, 48:1005–1013.

Tables

Table 1: Conditions of simulations for type I and type II error rates: the number of sampling sites, the testing method used (parametric of permutations), and the number of species simulated on the sampling sites. 10000 trials were performed for each set of conditions for a total of 440 000 simulations.

Number of sites (N)	Test	Number of species (M)			
10	Parametric	1	2	3	5
10	Permutations	1	2	3	5
25	Parametric	1	3	5	10
25	Permutations	1	3	5	10
50	Parametric	1	5	10	20
50	Permutations	1	5	10	20
100	Parametric	1	10	20	50
100	Permutations	1	10	20	50
250	Parametric	1	20	50	100
500	Parametric	1	50	100	250
1 000	Parametric	1	100	250	500

Table 2: Statistically significant components of the multivariate spatial codependence between fish assemblages (Hellinger-transformed counts) and descriptors of water quality; permutation tests.

Scale	Descriptor	ϕ_{ν_1,ν_2}	ν_1	ν_2	P
MEM_1	$flow$	2434.3	27	27	0.005
MEM_4	BOD	30.67	27	26	0.01
MEM_3	$[NH_4^+]$	27.78	27	25	0.01
MEM_2	$[O_2]$	42.85	27	24	0.01

Table 3: Statistically significant components of the multivariate spatial codependence between Oribatid mite community structure (Hellinger-transformed counts) and microhabitat descriptors ($WaterCont$: water content of the peat; ($Shrub : Many$): dummy variable representing the highest of three ordered classes of shrub cover; ($Subs : Sphagnum1$): dummy variable representing peat containing *Sphagnum rubellum* with some *S. magellacinum*; ($Topo$: *hummock*): dummy variable representing a raised micro-topography. Permutation tests.

Scale	Descriptor	ϕ_{ν_1,ν_2}	ν_1	ν_2	P
MEM_1	$WaterCont$	1785.1	35	68	0.005
MEM_4	$Shrub : Many$	324.4	35	67	0.005
MEM_2	$Subs : Sphagnum1$	51.15	35	66	0.01
MEM_3	$Topo : hummock$	67.52	35	65	0.01

Figures

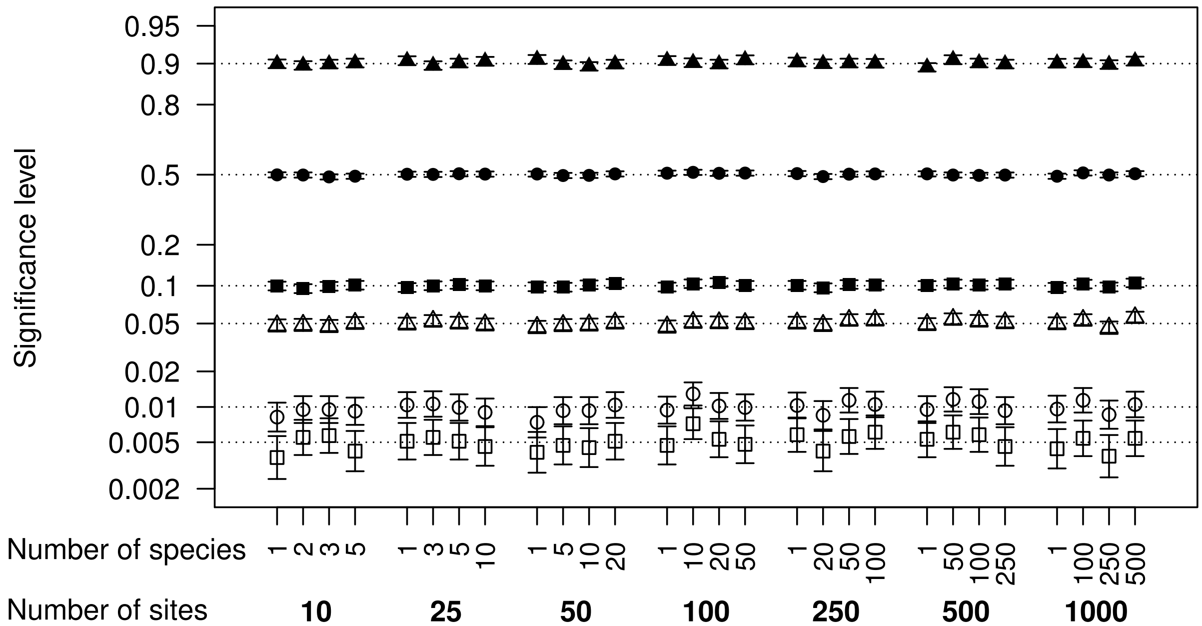

Figure 1: Simulation results, type I error rates, parametric test. Estimated mean rejection rates (with 95% confidence limits) for the null hypothesis of no codependence between response variables and a single explanatory variable, for different sample sizes. Abscissa: number of sites and response variables (called species). Simulated data were normally-distributed. Rates are shown for six different α significance levels, namely, 0.9 (▲), 0.5 (●), 0.1 (■), 0.05 (△), 0.01 (○), and 0.005 (□). 10 000 data set were simulated for each result shown.

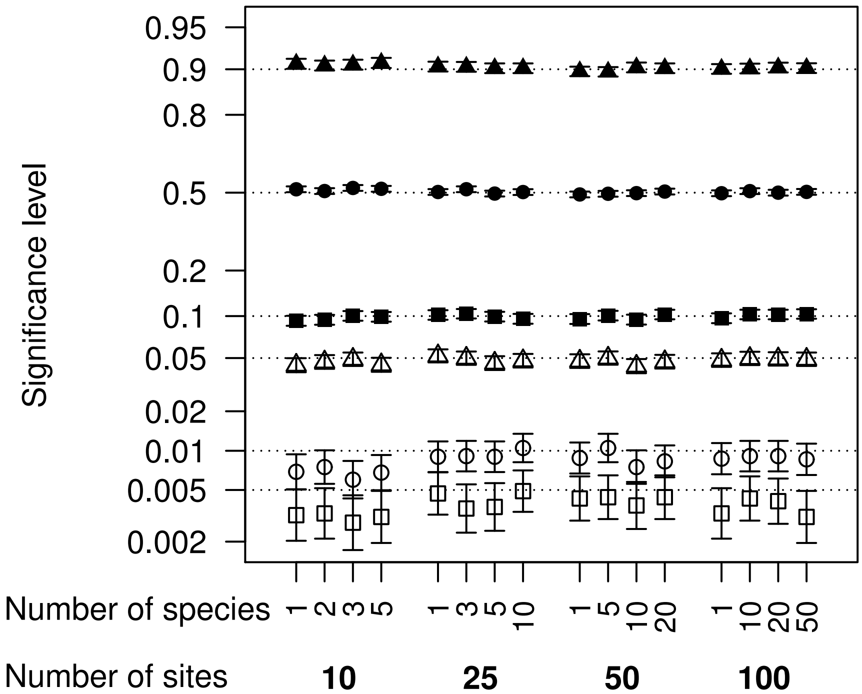

Figure 2: Simulation results, type I error rates, permutation test. Estimated mean rejection rates (with 95% confidence limits) for the null hypothesis of no codependence between response variables and a single explanatory variable, for different sample sizes. Abscissa: number of sites and species. The simulated species data were over-dispersed counts obtained by generating random normal deviates with a mean of 0 and a standard deviation of 1.5, exponentially-transforming them, and truncating them to the lowest integer. Rates are shown for six different α significance levels, namely, 0.9 (▲), 0.5 (●), 0.1 (■), 0.05 (△), 0.01 (○), and 0.005 (□). 10 000 data set were simulated for each result shown.

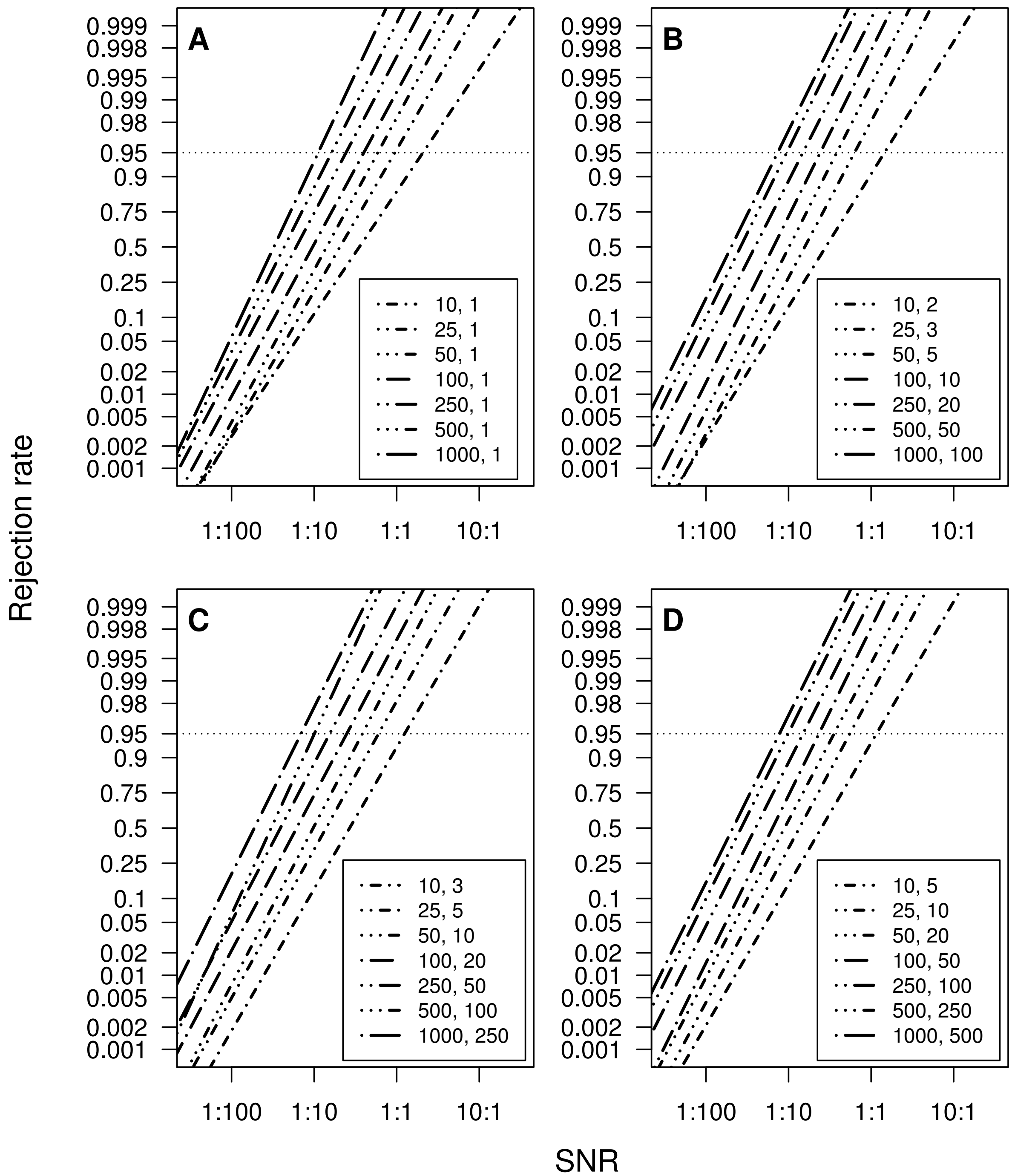

Figure 3: Parametric test: estimated statistical power (rejection rate) as a function of the signal-to-noise ratio (SNR) under an α significance level of 0.05 and different sample sizes (box legend: number of sites N, number of response variables M) represented by the different line types. Simulations encompass the single species (univariate) case (A) as well as cases with small (B), large (C), and very large (D) numbers of response variables (called species) with respect to the number of sampling sites. The 95% confidence limits of the lines were not shown because they were narrower than their line widths.

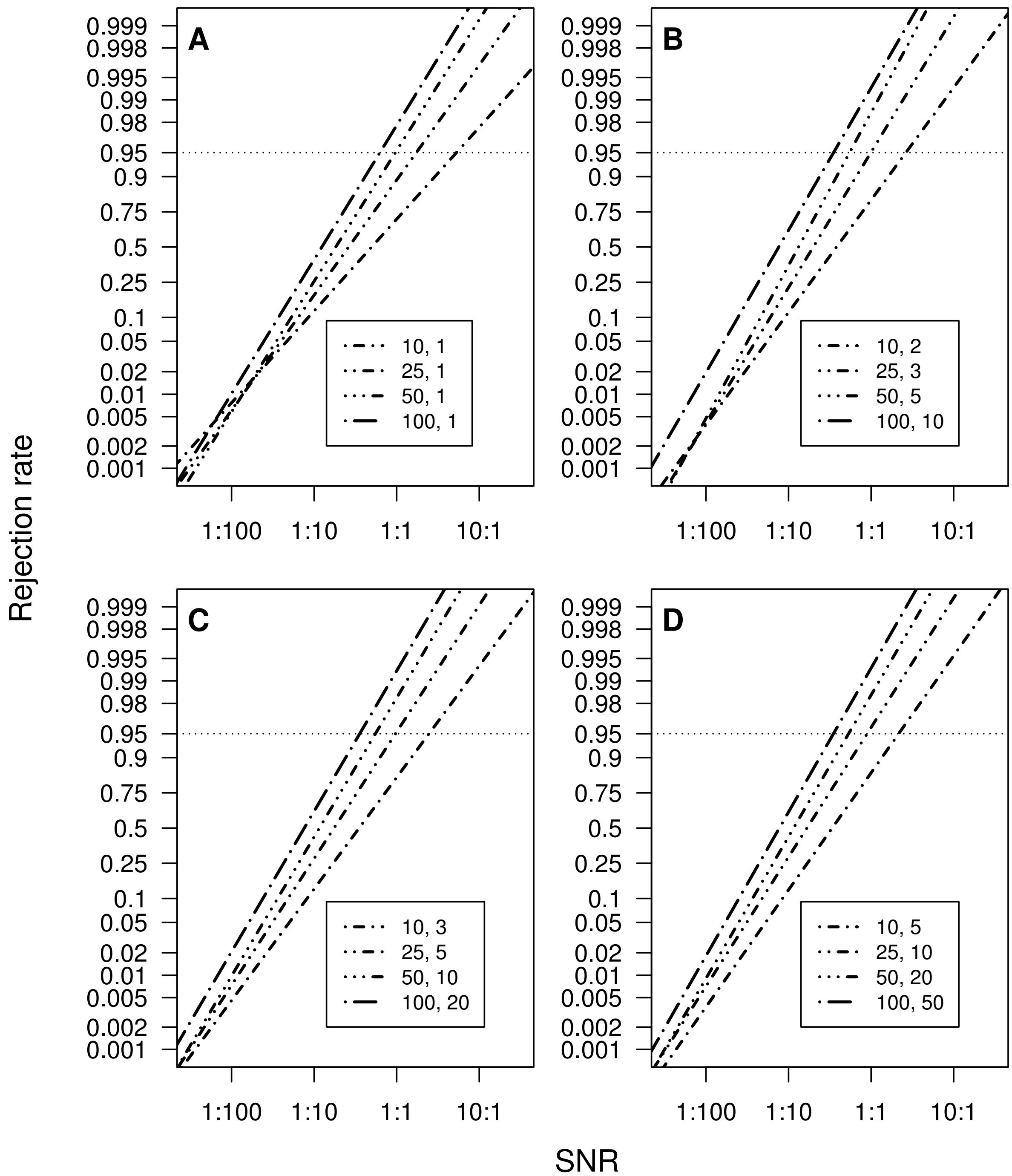

Figure 4: Permutation test: estimated statistical power (rejection rate) as a function of the signal-to-noise ratio (SNR) under an α significance level of 0.05 and different sample sizes (box legend: number of sites N, number of species M) represented by the different line types. Simulations encompass the single species (univariate) case (A) as well as cases with small (B), large (C), and very large (D) numbers of species with respect to the number of sampling sites. The 95% confidence limits of the lines were not shown because they were narrower than their widths.

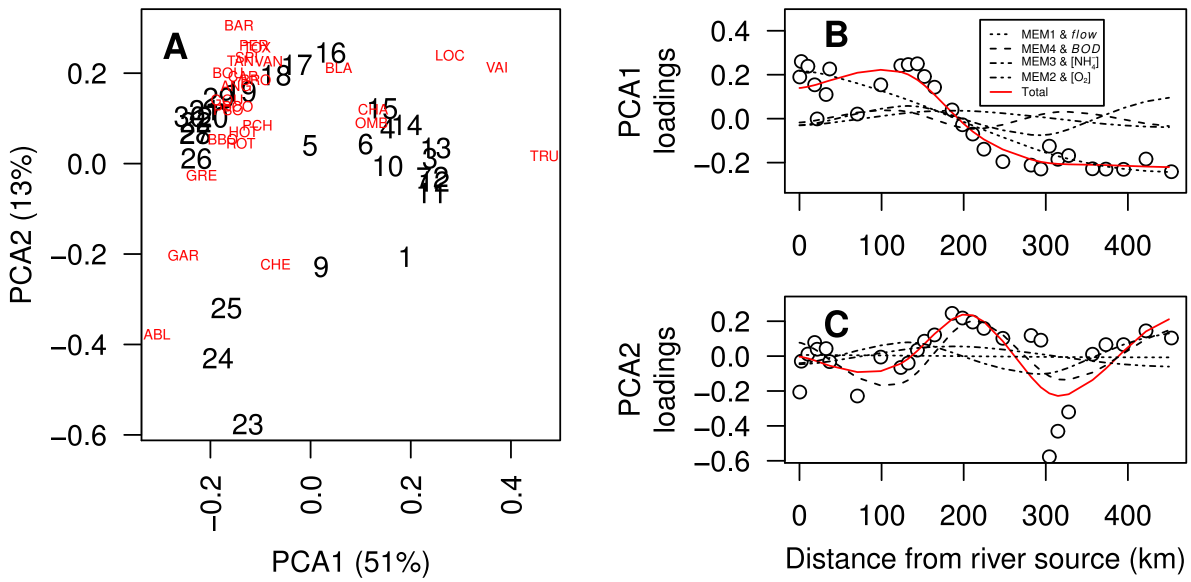

Figure 5: The statistically-significant spatial components of the codependence between fish community structure (represented as the first two principal components of community variation, A; numbers refer to the sites, in order from headwaters (site 1) to river mouth (site 30) whereas red labels refer to the species (see Verneaux, 1973, for corresponding Latin names), and four descriptors of water quality in Doubs River (France), namely $flow$: river discharge, BOD: biological oxygen demand, $\left[\mathrm{NH}_4^+\right]$: ammonium concentration, $[\mathrm{O}_2]$: dissolved oxygen. Panels B and C: contributions of the significant MEM spatial components (and their total effect) to the first and second principal components of fish community variation, respectively.

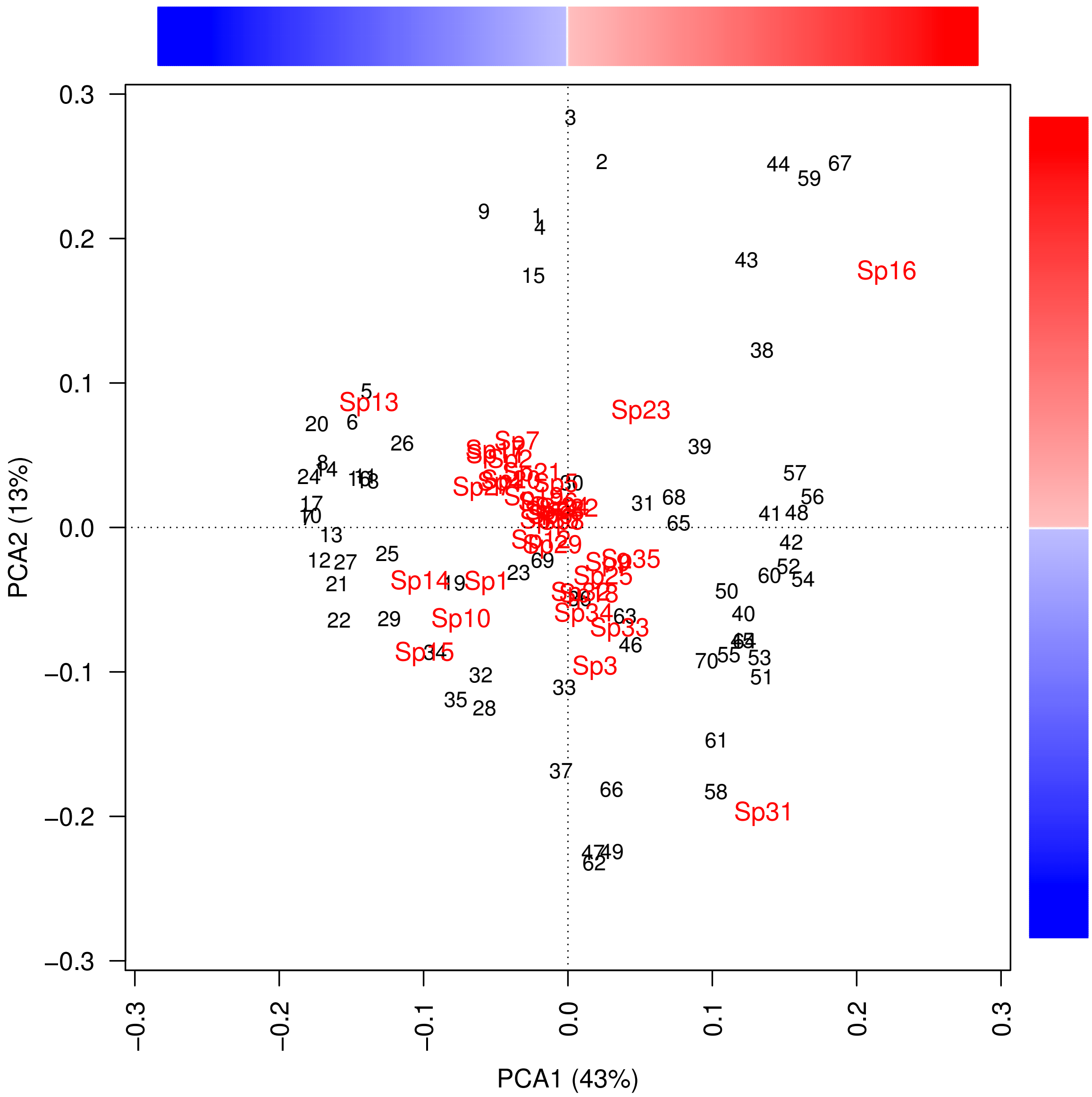

Figure 6: The first two principal components of the Oribatid mite community structure. The 70 sites are labelled using numbers whereas the 35 morpho-species are labelled as Sp1 to Sp35 (in red). The colour scale represents values on the principal components; it is used in Figs. 7–8.

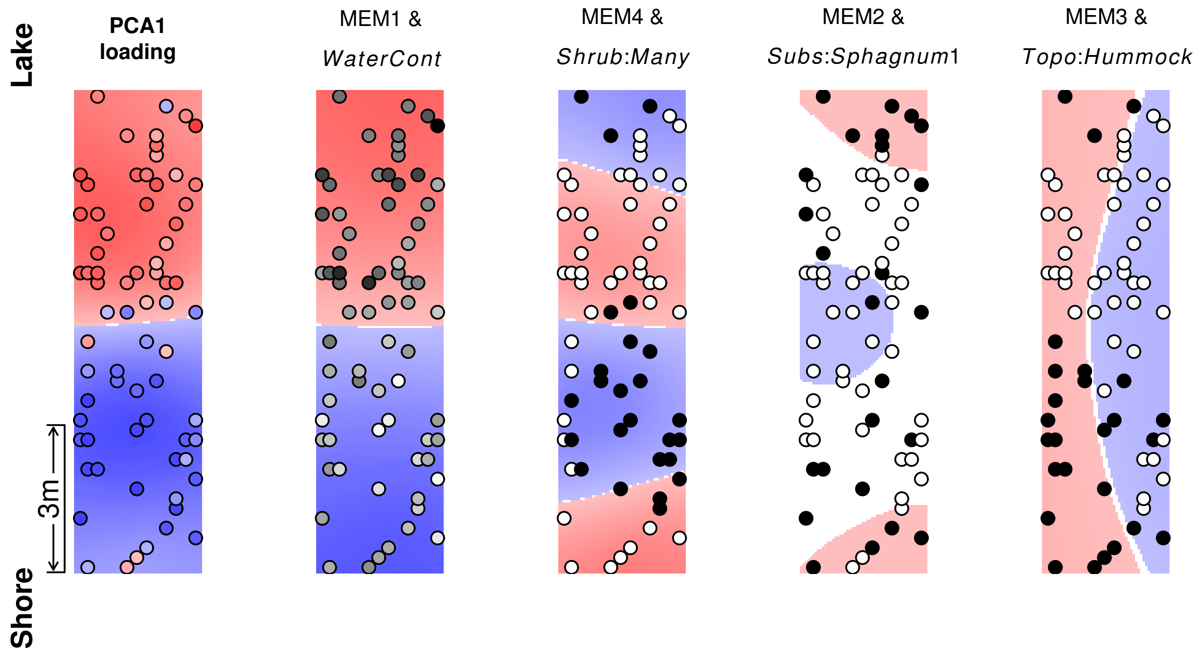

Figure 7: Geographic map of the mite data showing the statistically-significant spatial components of the codependence between the mite community structure and the environmental variables to which it is associated at certain spatial scales, as found by the analysis. The left panel shows the 70 sites with colours corresponding to their positions along PCA1 (Fig. 6). The following panels show the 70 sites again with symbols shaded according to the value of the environmental variable shown at the top of the map, and background colours corresponding to the positive and negative portions of the MEM giving the scale of the mite-environment codependence.

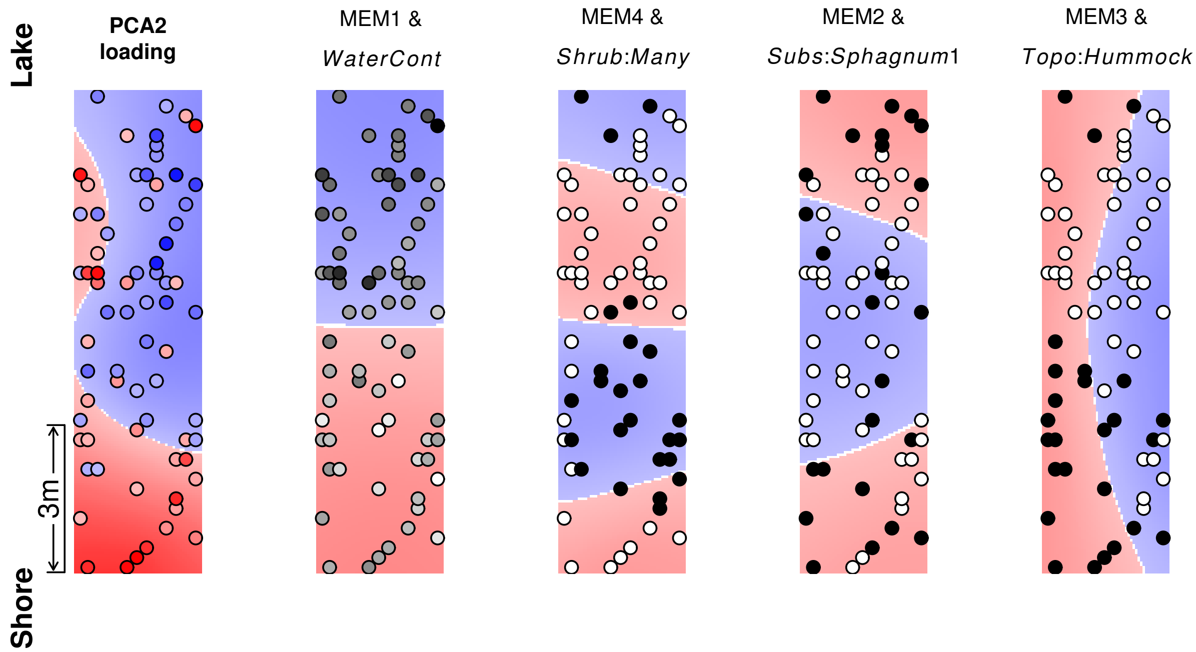

Figure 8: Geographic map of the mite data showing the statistically-significant spatial components of the codependence between the mite community structure and the environmental variables to which it is associated at certain spatial scales, as found by the analysis. The left panel shows the 70 sites with colours corresponding to their positions along PCA2 (Fig. 6). The following panels show the 70 sites again with symbols shaded according to the value of the environmental variable shown at the top of the map, and background colours corresponding to the positive and negative portions of the MEM giving the scale of the mite-environment codependence.