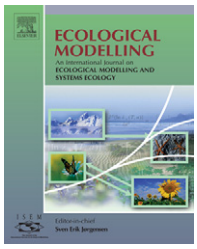
ECOLOGICAL MODELLING 215 (2008) 325–336

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/ecolmodel



Modelling directional spatial processes in ecological data

F. Guillaume Blanchet[a,b,*], *Pierre Legendre*[a], *Daniel Borcard*[a]

[a] *Département de sciences biologiques, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Québec, Canada H3C 3J7*
[b] *Department of Renewable Resources, University of Alberta, 751 General Service Building, Edmonton, Alberta, Canada T6G 2H1*

ARTICLE INFO

Article history:
Received 10 August 2007
Received in revised form
20 March 2008
Accepted 4 April 2008
Published on line 27 May 2008

Keywords:
Directional spatial process
Geographic eigenfunctions
Mastigouche Reserve
Moran's eigenvector maps (MEM)
Salvelinus fontinalis
Spatial analysis
Spatial autocorrelation
Spatial model

ABSTRACT

Distributions of species, animals or plants, terrestrial or aquatic, are influenced by numerous factors such as physical and biogeographical gradients. Dominant wind and current directions cause the appearance of gradients in physical conditions whereas biogeographical gradients can be the result of historical events (e.g. glaciations). No spatial modelling technique has been developed to this day that considers the direction of an asymmetric process controlling species distributions along a gradient or network. This paper presents a new method that can model species spatial distributions generated by a hypothesized asymmetric, directional physical process. This method is an eigenfunction-based spatial filtering technique that offers as much flexibility as the Moran's eigenvector maps (MEM) framework; it is called asymmetric eigenvector maps (AEM) modelling. Information needed to construct eigenfunctions through the AEM framework are the spatial coordinates of the sampling or experimental sites, a connexion diagram linking the sites to one another, prior information about the direction of the hypothesized asymmetric process influencing the response variable(s), and optionally, weights attached to the edges (links). To illustrate how this new method works, AEM is compared to MEM analysis through simulations and in the analysis of an ecological example where a known asymmetric forcing is present. The ecological example reanalyses the dietary habits of brook trout (*Salvelinus fontinalis*) sampled in 42 lakes of the Mastigouche Reserve, Québec.

1. Introduction

It is well known that spatial distributions of species are influenced by environmental gradients (Huston, 1996). Since the article of Legendre and Fortin (1989), the importance of spatial structures has been well understood by ecologists. This has led to a number of methodological developments to study spatial patterns in ecology. Methods devised in other domains have also been applied to ecology. For example, geostatistical tools have been, and still are, used to investigate spatial relationships in an ecological perspective; Peterson et al. (2007) is a recent example of the use of geostatistics in river modelling. Legendre (1990) proposed to use polynomials of the geographic coordinates of the sites to represent spatial relationships in models aimed at explaining species variation. More recently, the development of principal coordinates of neighbour matrices (PCNM) (Borcard and Legendre, 2002; Borcard et al., 2004; Legendre and Borcard, 2006) has provided a new and more powerful way for studying spatial variation. It has also significantly enhanced the proportion of variation explained by spatial models. Dray et al. (2006) developed the framework of Moran's eigenvector maps (MEM), which is a generalization of the PCNM approach. Griffith and Peres-Neto (2006) unified the

* *Corresponding author at*: Department of Renewable Resources, University of Alberta, 751 General Service Building, Edmonton, Alberta, Canada T6G 2H1. Tel.: +1 780 492 8670.
E-mail addresses: gblanche@ualberta.ca (F.G. Blanchet), Pierre.Legendre@umontreal.ca (P. Legendre), Daniel.Borcard@umontreal.ca (D. Borcard).
0304-3800/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2008.04.001

PCNM, MEM, and spatial filtering methods (Griffith, 2000) into a family called eigenfunction-based spatial analysis. Borcard et al. (1992) showed through variation partitioning that spatial relationships and environment can explain both separate and common variation of the distributions of species. To this day, however, no methodological development has shown how to model the influence of asymmetric, directional process on species distributions or other response variables of interest.

At broad or fine scales, the spatial distribution of species is often structured by abiotic and/or biotic gradient(s). We propose that gradients influencing species spatial distributions can be studied via spatial variables (eigenfunctions) that represent directional spatial processes. Dray et al. (2006) deplored the absence of methods capable of modelling asymmetric directional spatial processes; the present paper fills that gap. Here, a new framework is presented, which is also part of the eigenfunction-based spatial filtering framework, with the added feature that it considers space in an asymmetric way. Variables created via this framework will be called asymmetric eigenvector maps (AEM). This method was created for situations where a hypothesized asymmetric, directional spatial process influences the species distribution at scales ranging from fine to broad (e.g. the directional effects of a river network, or of currents in a sea, river, stream, or fluvial lake, on species distributions). Since the AEM framework creates spatial variables corresponding to an asymmetric, directional process, these variables can be used to model the spatial structure of any set of response variables, e.g. single-species population data, multi-species community, meta-population, or meta-community, influenced by an asymmetric spatial process. A process is a phenomenon organized along space and/or time. We define a directional spatial process as a process that can be represented by directional arrows in geographic space. The structures resulting from directional processes are asymmetric. To test the functioning and limits of the new AEM method, simulations have been carried out in a two-dimensional spatial context.

2. Method

The Dray et al. (2006) MEM method consists in the diagonalization of a spatial weighting matrix (**W**). Matrix **W** is a resemblance matrix that can be constructed through the Hadamard product between two previously computed resemblance matrices: a connectivity matrix showing which sites are linked to one another by connexions, and a weighting matrix which gives the weight associated to each pair of sites. As developed by Dray et al. (2006), no direction can be imposed on the created MEM spatial variables because the framework is based on resemblance matrices that do not account for asymmetry.

The simplest form of data leading to AEM construction is a tree-like structure, like a river network. The relationships among the sampling sites can be written as described by Legendre and Legendre (1998, section 1.5.7): for each site, the river links (called "edges" hereafter, using the vocabulary of graph theory) located upstream from that site in the river network and considered to be influencing it receive the code "1" in a sites-by-edges table **E**; all other edges receive the code "0". The new development to this coding method, proposed here, is to transform table **E** into eigenfunctions. This can be done in three computationally different but otherwise equivalent ways:

1. Compute a principal component analysis (PCA) of table **E** and use matrix **F** of the principal components as the new matrix of explanatory variables. PCA scaling (type 1 or type 2) does not matter for the present application.
2. Alternately, compute a singular value decomposition (SVD) of the column-centred table **E**, called $\mathbf{E}_c$. Decompose $\mathbf{E}_c$ by SVD into $\mathbf{U}\,\mathbf{D}\,\mathbf{V}'$; **U** and **V** are column-orthonormal matrices and $\mathbf{V}'$ means **V** transposed. Use the left-hand column-orthonormal matrix **U**, resulting from the decomposition, as the new matrix of explanatory variables; **U** is linearly related to matrix **F** containing the principal components obtained by PCA and, for the present application, is equivalent to it.
3. A third alternative is to compute an Euclidean distance matrix among the rows of table **E**. A principal coordinate analysis (PCoA) of that distance matrix produces the same matrix **F** as obtained above by PCA.

Contrary to PCNM and MEM, AEM analysis produces no negative eigenvalues because a covariance matrix is a positive semidefinite matrix; hence, all PCA eigenvalues are positive or null (Legendre and Legendre, 1998, p. 138). The construction of AEM is presented in more detail in the next paragraphs and in Fig. 1. AEM eigenfunctions can be constructed from a river network (example developed above) or from other types of directional connexion networks. An ecological example presented later in this article to illustrate the use of AEM, will start from a set of lakes in a single hydrographic network. The analysis will attempt to explain the variation in brook trout (*Salvelinus fontinalis*) gut contents in 42 lakes of the Mastigouche Reserve, Québec.

AEM are based on a directional connexion network. Connexion networks can be constructed to correspond to hydrological (example above; Fig. 5) or other dynamic information available about the sampling units. In the absence of a precise dynamic model, they can be constructed using graph theory (e.g. Berge, 1958; Barthélemy and Guénoche, 1988).

A general type of connexion network for a regular sampling grid is shown in Fig. 1b. To impose directionality on the diagram and create asymmetric spatial variables, an imaginary site (site 0 in Fig. 1b) is added upstream of the sampling area. This fictitious site is connected to the uppermost true site(s) if, as in this example, the process influence is assumed to come from upstream. It is connected to the lowermost sites if the influence is hypothesized to come from downstream; that will be the case in the lake example presented later in this paper. In Fig. 1b, there are five sites that are equal in being the most upstream ones; site 0 is thus connected to all these sites (dashed lines). To quantify the connexions (edges) between the sites and construct matrix **E**, a method originally proposed for phylogenetic reconstruction by Kludge and Farris (1969: binary coding of a transformation series) will be used. Sites (rows of table **E**) and edges (columns) are numbered; alternatively, they can be given names. In the fictitious example, which involves a downstream process, each site is characterized by all the upstream edges connecting the site of interest to site 0,

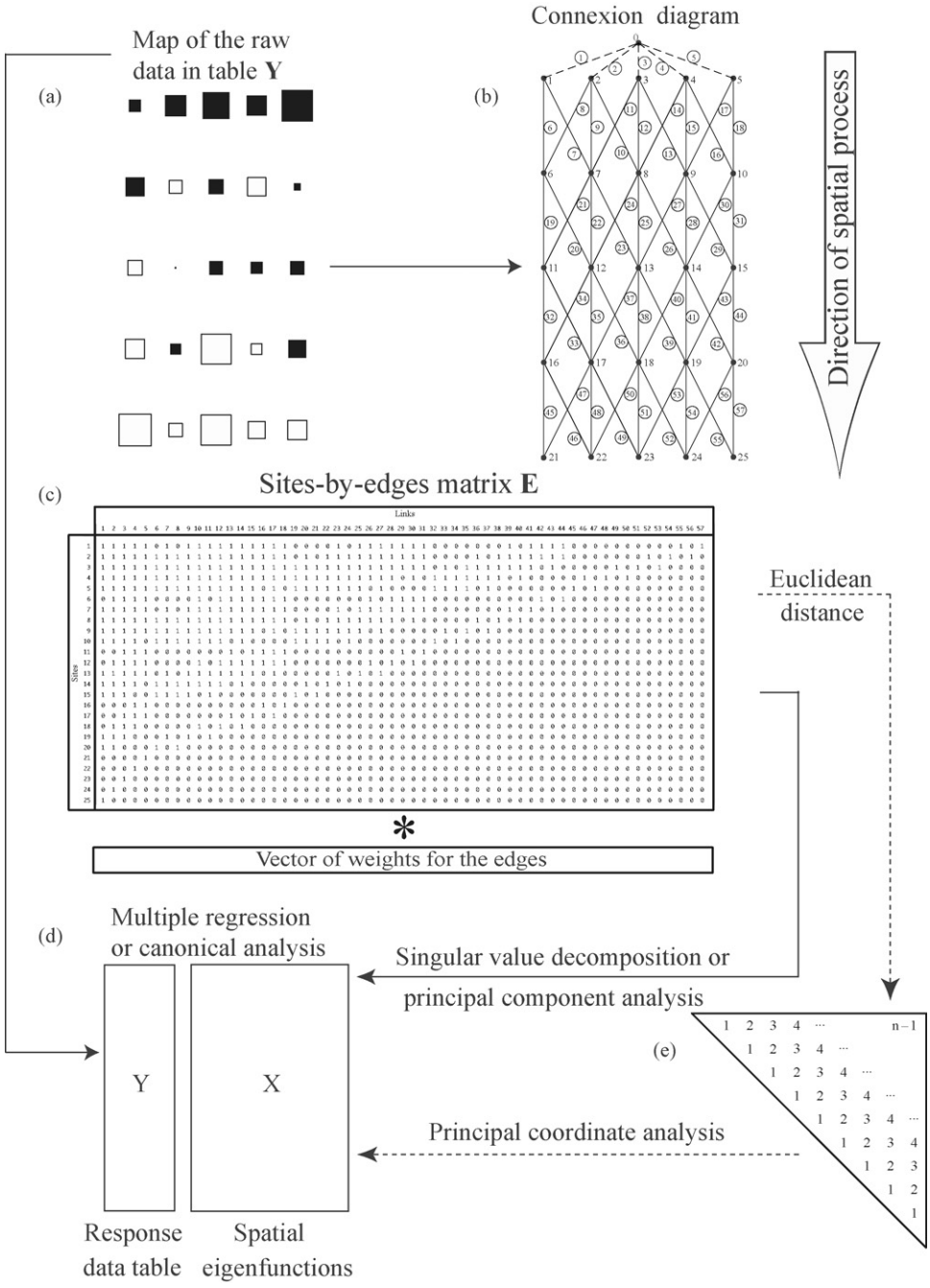


Fig. 1 – Schematic representation of AEM analysis using a fictive example. (a) Data values are represented by bubbles (empty = negative, full = positive values). (b) Sites are linked by a connexion diagram, which in turn will be used to construct the sites-by-edges matrix E (c). Weights can be attributed to the edges (columns) of this matrix, representing the difficulty of effect transmission between nodes (vector underneath the sites-by-edges matrix). (d) Descriptors (AEM variables, matrix X) are obtained by calculating the left-hand matrix of eigenvectors of SVD, or the matrix of principal components (site scores) of PCA. AEM variables (matrix X) can also be obtained through the calculation of an Euclidean distance matrix followed by the computation of eigenvectors via principal coordinate analysis (PCoA).

directly or indirectly. The sites-by-edges table **E** is filled with 0's and 1's representing the absence or presence of the various edges linking each site to site 0 (Fig. 1c). It is to be noted that site 0 is not present in this matrix because it is not influenced by any edge; if present, this site would add an unnecessary line to the matrix giving no additional information.

Weights can be added to the sites-by-edges matrix by multiplying a vector of weights to table **E**′ (Fig. 1c) (Ronquist, 1996). Weights can be given based on various types of known information, e.g. the lengths of the edges, which represent distance, or more generally the difficulty of transfer of the process effect between nodes of the network.

The eigenfunctions created with this method are orthogonal variables, as is the case for the eigenfunctions created by the PCNM and MEM methods. This is because they are eigenvectors of a symmetric matrix. Computation through the calculation of a distance matrix followed by principal coordinate analysis (computation method 3 above), as well as the possibility to add weights to the links, show the closeness of the AEM (the present paper) and MEM methods (Dray et al., 2006).

The AEM framework sometimes generates eigenfunctions that have the same weight (i.e., two or more eigenvectors have the same eigenvalue). This can also occur in the MEM framework. This will need further investigation to better understand under what circumstances these are generated and how to handle and interpret them.

3. Simulation study

We carried out a range of simulations to better understand the behaviour of AEM eigenfunctions in different situations. AEM eigenfunctions were first tested for type I error. For power evaluation, they were compared to MEM eigenfunctions in the presence of asymmetric generating processes, for different types of spatial structures, using the proportions of variance explained as criterion.

Simulations were first used to estimate the type I error of AEM analysis. Two sets of simulations with a hundred points were produced, representing opposite extremes of the AEM framework: (1) the points were regularly distributed on a ten-by-ten grid (see Fig. 2a for the connexion network), no weights were attached to the edges; (2) the points were irregularly distributed on the map (see Fig. 2c for the connexion network) and the edges were weighted by the inverse of the distances. Following Manly (1997) and Anderson and Legendre (1999), the response variables contained values drawn at random from four distributions: normal, uniform, exponential, and exponential cubed. The relationship between the random response variables and the AEM eigenfunctions was tested at the 5% significance level. Because there are $n-1$ eigenfunctions created by the AEM procedure, where n is the number of points, one

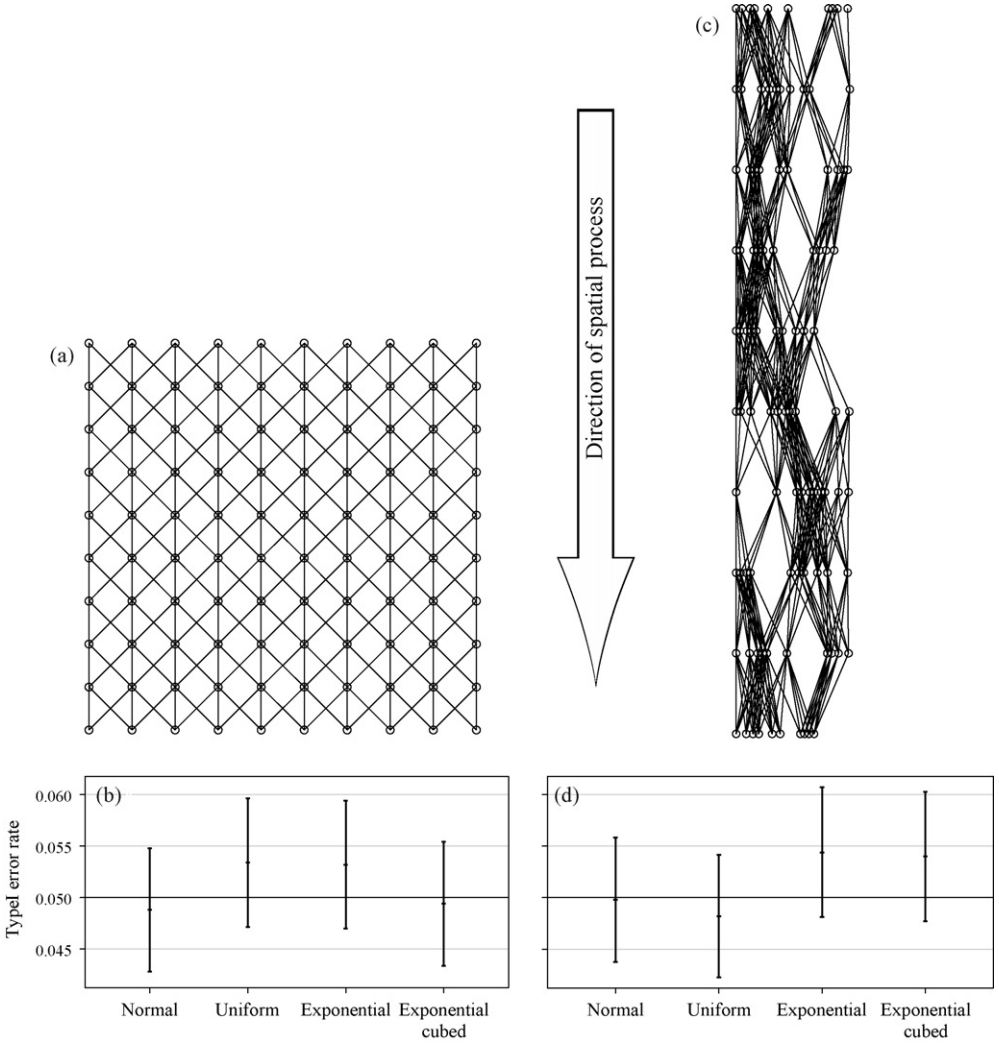
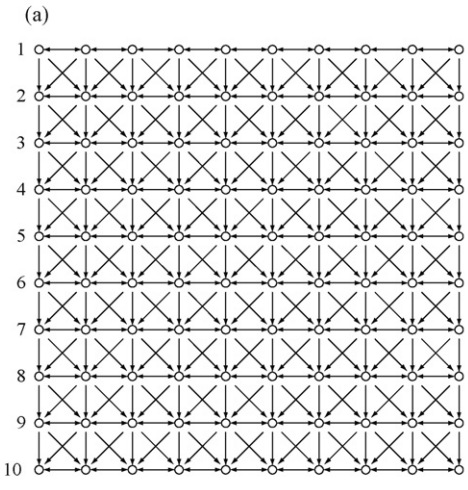


Fig. 2 – Type I error of AEM analysis (b, d) for sampling points and connexion diagrams shown in (a) and (c). No weights were used in (a), whereas the inverses of the distances were used as weights in (c). The large arrow represents the direction of the asymmetric process considered in (a) and (c). Response values were randomly selected for each point from four different distributions. Each run consisted of 5000 independent simulations. The errors bars in (b) and (d) represent 95% confidence intervals on the rejection levels.

cannot carry out a test of significance using all eigenfunctions. Following Blanchet et al. (in press), the AEM eigenfunctions were divided in two groups depending on the value of the associated Moran's I coefficients. The Moran's I coefficients were computed using only the direct links between sites. The first group contained the eigenfunctions with Moran's I values higher than the expected value; these were positively autocorrelated. The second group, which contained the eigenfunctions with Moran's I values lower then the expected value, were negatively autocorrelated. The two sets of eigenfunctions were tested separately for significance (permutation test, 999 random permutations). Since two p-values were calculated, they were subjected to a Sidak (1967) correction. If at least one of the two p-values was significant after correction, the relationship was considered to be significant. Fig. 2b and d present the results for the two series of simulations. Each reported value is the result of 5000 independent simulations. In all cases, the number of significant results was very close to the 5% significance level. These results show that the AEM method has a correct level of type I error in the two examined situations, and this for the four types of error distributions.

For the power analysis, simulations were carried out to see how well various subsets of the AEM eigenfunctions react in the presence of gradients, when compared to MEM eigenfunctions. These simulations were done on a ten-by-ten regular grid (Fig. 3a); thus $n = 100$. Eight different structures were used to generate the data in these simulations (Fig. 3b). The eight structures were generated in such a way that in each pair of structures (S1–S2, S3–S4, S5–S6, and S7–S8), one represents a symmetric gradient from row 1 to row 10 whereas the other is an asymmetric gradient. The odd-numbered structures are the asymmetric gradients. These structures were each tested with three univariate and one multivariate response data sets. In the three univariate situations, a random normal error with a mean of 0 and standard deviation (S.D.) of 1, 2 and 3 was added to the structure. Standard deviations larger than 3 were not considered because in all situations except S1 and S2, the basic structure of the data did not have "steps" higher than 3. For the multivariate situation, 10 response variables were generated, 5 containing structure and noise (random error) and 5 containing noise only. The error values were drawn at random from a normal distribution with mean 0; the standard deviation was randomly drawn, for each simulation, from a uniform distribution between 1 and 3. For each set, one thousand simulations were carried out.

When weights are placed on the edges, both the AEM and MEM frameworks can create an infinite number of different spatial variables for a set of sites. We decided to include 21 different combinations of functions and weights in our comparisons; thus 21 different sets of spatial variables (eigenfunctions) were created for each framework. The connexion diagram used was the same in all situations to allow appropriate comparisons (Fig. 3a). Note that not all edges have the same meaning in the AEM network. The horizontal edges represent the lateral influence of a site on its nearest neighbours, while the vertical and oblique edges represent the hypothesized asymmetric process. Because of this difference, each set of edges was considered independently. When the sites-by-edges matrix was constructed for the connexion diagram illustrated in Fig. 3a, all these edges became the columns of the sites-by-edges matrix **E** of Fig. 1c. In the creation of this sites-by-edges matrix, the horizontal edges did not influence the construction of the columns that were built from the vertical or oblique edges, and vice versa. Weights were given to the edges based on the concave-down ($f_1 = 1 - d_{ij}/\max(d_{ij})^{\alpha}$) and concave-up ($f_2 = 1/d_{ij}^{\alpha}$) distance functions, as in Dray et al. (2006). Ten different values of the exponent α, from 1 to 10, were used for each function. For both the AEM and MEM frameworks, combination 21 consisted in a series of spatial variables constructed with uniform weights of 1 for all edges; this combination never came out in the results as the one producing the highest explained variance. Each weighted sites-by-edges



(b)

	S1	S2	S3	S4	S5	S6	S7	S8
1	6	6	6	6	9	9	12	12
2	6	6	6	6	9	9	12	12
3	0	6	3	6	6	9	9	9
4	0	6	3	6	6	6	6	9
5	0	6	0	3	3	6	3	6
6	0	0	0	3	3	3	3	6
7	0	0	0	0	0	3	3	3
8	0	0	0	0	0	0	0	3
9	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0

Fig. 3 – (a) Connexion diagram used to create AEM and MEM eigenfunctions. Arrows represent directions of influence of sites on each other; these directions were taken into account during the construction of AEM eigenfunctions, but not for MEM eigenfunctions. The rows of data points are numbered. (b) Eight basic structures (S1 to S8, columns) used to generate the response variables. The numbers are the values added to all points on each line (1–10) of the diagram in (a), prior to adding random normal noise.

Table 1 – Weighting function (f_1, f_2) and exponent α giving the highest explained variance when modelling each structure in each set of simulations, with AEM or MEM

Response	Structure (S1 to S8)	AEM		MEM	
		Weighting function	Exponent α	Weighting function	Exponent α
Univariate S.D. = 1	1	f_1	4	f_2	9
	2	f_1	3	f_2	5
	3	f_2	8	f_2	8
	4	f_1	4	f_2	5
	5	f_1	10	f_2	10
	6	f_1	5	f_2	2
	7	f_1	10	f_2	9
	8	f_2	6	f_2	5
Univariate S.D. = 2	1	f_1	10	f_2	8
	2	f_2	2	f_2	5
	3	f_2	3	f_2	9
	4	f_2	3	f_2	5
	5	f_1	9	f_2	9
	6	f_1	2	f_2	2
	7	f_2	9	f_2	9
	8	f_1	4	f_2	7
Univariate S.D. = 3	1	f_2	8	f_2	9
	2	f_1	4	f_2	8
	3	f_2	5	f_2	9
	4	f_1	6	f_2	5
	5	f_1	9	f_2	10
	6	f_1	4	f_2	4
	7	f_2	4	f_2	9
	8	f_2	6	f_2	6
Multivariate	1	f_1	2	f_2	8
	2	f_1	8	f_2	7
	3	f_2	3	f_2	10
	4	f_2	3	f_2	8
	5	f_1	7	f_2	10
	6	f_2	7	f_2	3
	7	f_1	1	f_2	10
	8	f_2	10	f_2	5

The chosen combination of weighting function and exponent, in each case (2 weighting functions and 10 exponents α), was the one that produced the highest value of (R^2_a). The same response variables were used in the AEM and MEM simulations. S.D. = standard deviation.

matrix was then used as the table of explanatory variables for the simulated data. Because there are always $(n-1)$ AEM variables and often also $(n-1)$ MEM variables, the same procedure used to test the type I error of the AEM eigenfunctions was used here to test the significance of each set of spatial variables. The eigenfunctions were divided in two groups, positively and negatively autocorrelated, using the eigenvalues associated with the eigenfunctions; Dray et al. (2006) have shown that there is a direct correlation between Moran's I and the eigenvalues produced in the MEM framework. The test used for the univariate simulations is a parametric test in multiple regression; that test was appropriate because the error was normally distributed by construction. In the multivariate simulations, the generated response data were analyzed as a function of the AEM and MEM eigenfunctions by canonical redundancy analysis (RDA), followed by a permutation test produced by the "anova.cca" function of the "vegan" package (Oksanen et al., 2007) in the R statistical language (R Development Core Team, 2007). That procedure allows the function to propose a statistical decision (reject H_0 or not) after 99 to 499 random permutations by steps of 100. For each particular type of data structure (S1 to S8), the AEM and MEM results that are compared (1000 simulations) are those corresponding to the eigenfunctions, obtained from a given weighting function (f_1, f_2) and exponent, that explained, on average, the largest amount of variance (R^2_a) of the response data, while still being significant at the 5% level. These choices are listed in Table 1. The results for the univariate and multivariate simulations are presented in Fig. 4.

Due to the inherent structure of the simulated data, we were expecting to obtain better results with AEM only when the structure of the gradient was asymmetric (odd-numbered structures). Actually, the AEM variables turned out to reject the null hypothesis and identify a significant structure more often than MEM eigenfunctions in all situations, except for S1, S3 and S7 when S.D. was large (Fig. 4c), meaning that a lot of random noise was present in the data; then, the amount of explained variance (R^2_a) was roughly the same for AEM and MEM, the confidence intervals being superposed. This result surprised us because it showed that the AEM framework, though it creates variables that represent asymmetric processes by construction, is not only better suited than MEM

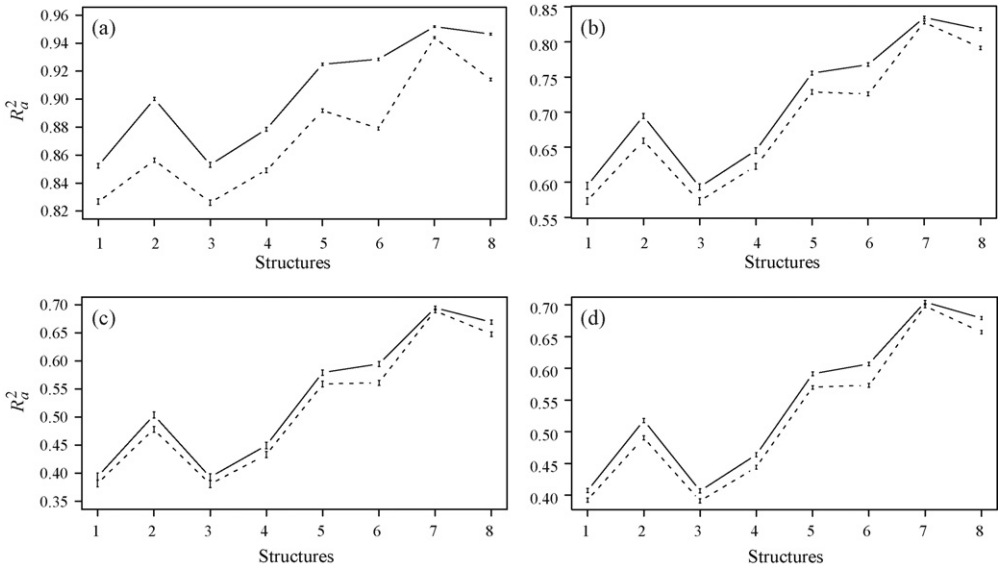


Fig. 4 – Variance explained (R^2_a) for the best set of AEM (full lines) and MEM (dashed lines) variables for each of the 8 structures described in Fig. 3b. Panels (a–c) present results of univariate simulations where the error term values were randomly drawn from a normal distribution with standard deviations of 1, 2, and 3, respectively. Panel (d) presents results of multivariate simulations where the error term values were randomly chosen from a normal distribution whose standard deviation was selected at random from a uniform distribution with a minimum of 1 and a maximum of 3. Vertical error bars represent 95% confident intervals on the rejection rates. Each run consists of 1000 independent simulations. Lines linking error bars were plotted to prevent confusion between the results of the AEM and MEM analyses.

for asymmetric data, it is also equally or more appropriate than MEM variables in all gradient situations. AEM variables produced results roughly equivalent to those of MEM analysis only in the presence of abrupt changes in the gradient. S7 is a good example of such a situation. In more continuous cases, AEM analysis always performed better than MEM at identifying the gradient.

The weighting functions (f_1, f_2) that best modelled the simulated data were very different between the two frameworks. MEM variables created with function f_2 were always the best ones, but this was not always the case in AEM analysis. These results show that the difference in construction between the two methods can result in "best models" having very different weights. For real ecological data, if weights derived from the concave-down or concave-up functions are added to the edges, the interpretation given *a posteriori* to the weights can vary widely depending on the modelling framework (AEM or MEM). It is thus not advisable to add weights to the edges when the added weights do not come from some sort of prior hypothesis about the process having generated the data, despite the fact that adding weights can improve the empirical modelling ability of AEM eigenfunctions, as was shown by Dray et al. (2006) for MEM eigenfunctions.

When comparing the three sets of univariate simulations, the best MEM models were quite consistent between sets of simulations for each particular structure (S1 to S8): the correlation coefficients among the three sets of α parameter values are all near 0.90. This is not the case for AEM analysis, where the weighting function (f_1, f_2) and the α parameter value for the best model may change between sets of simulations. To deepen the investigation, we compared the variance explained by AEM models (R^2_a), on average, across each set of 1000 simulations. The means of the R^2_a statistics were very similar for different weights α; often the best and second-best results diverged by less than 0.1%. Table 1 would thus be likely to be different after another series of simulation; the amounts of explained variance presented in Fig. 4 would, however, not be different. This is related to the construction of AEM variables when weights are added. The way weights are considered in the AEM framework makes the variables less sensitive to the differences among weights, compared to MEM analysis. The weights used in these simulations do not favour the AEM framework: the results show that different weights create spatial variables explaining almost identical amounts of variation in AEM analysis; this is not the case for MEM eigenfunctions.

4. Ecological illustration

To illustrate the application of AEM analysis to real ecological situations, we used data collected on 42 lakes of the Mastigouche Reserve, Québec, Canada (46°40′N, 73°20′W) and analyzed by Magnan et al. (1994). The dependent data matrix describes brook trout (*Salvelinus fontinalis*) diet composition in those lakes. In each lake, 20 stomachs were sampled during daytime by anglers in June 1989. Mean percent wet mass was recorded for nine functional prey categories: zoobenthos, amphipods, zooplankton, dipteran pupae, aquatic insects, terrestrial insects, prey-fish, leeches, and other prey. More detailed accounts of the data are presented in Lacasse and Magnan (1992) and Magnan et al. (1994). Fig. 5 presents a schematic map of the river network in the study area.

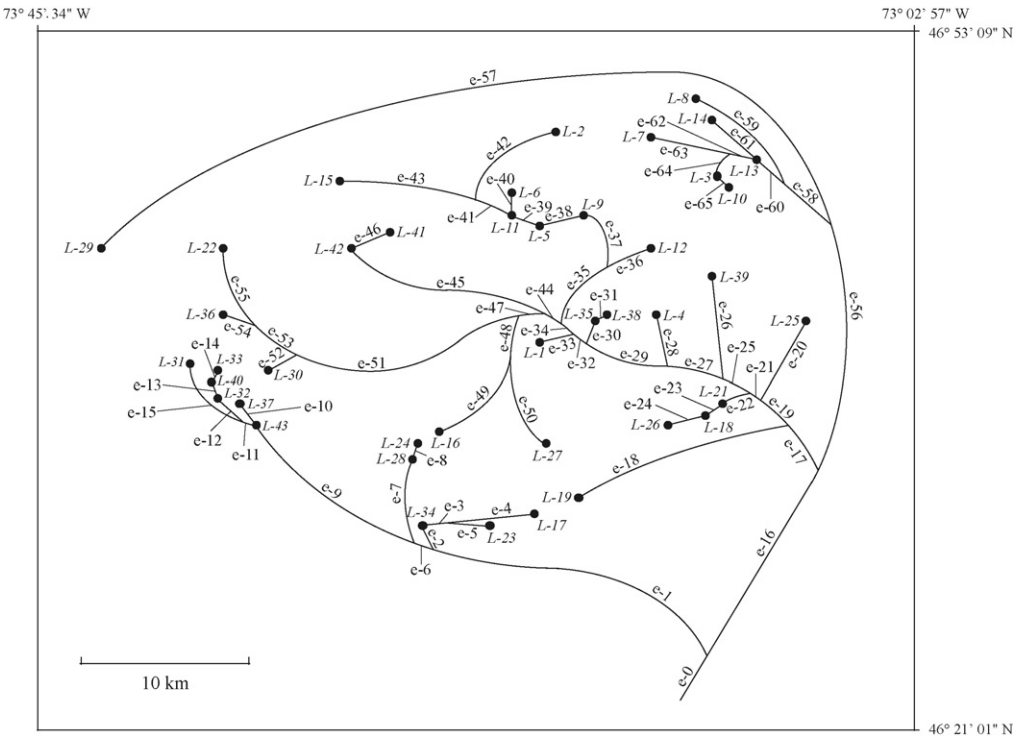


Fig. 5 – Schematic map of the river network in the Mastigouche Reserve. Lakes are numbered L-1 to L-43; there is no lake L-20. Edges are numbered e-1 to e-65; they are written to the sites-by-edges table E. Adapted from Magnan et al. (1994).

We compared AEM modelling to 6 other spatial modelling methods. The methods can be divided into three classes: those based on (1) lake geographic coordinates, (2) nodes of the river network, and (3) edges of the river network. Two analyses were done for type (1) data, a canonical correspondent analysis (CCA, ter Braak, 1986) using as explanatory variables a third-degree polynomial, and a canonical redundancy analysis (RDA, Rao, 1964) using principal coordinates of neighbour matrices (PCNM, Borcard and Legendre, 2002; Borcard et al., 2004). A CCA and an RDA, both based on nodes, were the methods used for type (2) data. The nodes used for the analyses are presented in Fig. 1 of the Magnan et al. (1994) paper. For type (3) data, we computed an RDA based on edges, an RDA based on Moran's eigenvector maps (MEM, Dray et al., 2006), and an RDA based on AEM spatial variables. Edges are labelled in Fig. 5. For each situation, a forward selection of spatial variables was

Table 2 – Comparison of spatial models of brook trout diet in 42 lakes, obtained from 7 different modelling methods

Modelling methods	No. spatial variables in full set	No. selected spatial variables	R^2	R^2_a
Method based on lake geographic coordinates				
CCA, 3rd deg. polynomial	9	4[a]	0.225	–
RDA, PCNM analysis	24	3[b]	0.257	0.199
Methods based on nodes of river network				
CCA, nodes	25	5[c]	0.356	–
RDA, nodes	25	4[d]	0.342	0.271
Methods based on edges of river network				
RDA, edges	65	9[e]	0.625	0.520
RDA, MEM analysis	41	11[f]	0.669	0.562
RDA, AEM analysis	41	13[g]	0.751	0.636

Forward selection was carried out using a cutoff level of $\alpha = 0.05$.

[a] Selected monomials: X, Y, Y^2, X^3.

[b] Selected PCNM variables computed from coordinates: 3, 4, 17.

[c] Selected nodes: 2, 9, 10, 12, 14. The nodes are shown in Fig. 1 of Magnan et al. (1994).

[d] Selected nodes: 10, 12, 14, 25.

[e] Selected edges: 21, 24, 27, 38, 46, 50, 52, 54, 58. Edges are shown in Fig. 5.

[f] Selected MEM variables computed from edges: 1, 3, 4, 6, 16, 17, 18, 20, 22, 27, 32.

[g] Selected AEM variables computed from edges: 1, 2, 3, 4, 6, 16, 18, 19, 22, 24, 25, 27, 29.

carried out using a cutoff level of $\alpha = 0.05$. For polynomial and node modelling, CCA was used instead of RDA to allow comparison with the results of Magnan et al. (1994); these authors used CCA on a subset of 37 lakes for which full environmental data were available. They used a cutoff level of $\alpha = 0.10$ in their forward selection in CCA. We used the full set of 42 lakes to obtain the results presented in Table 2. PCNM variables were constructed with a truncation distance equal to the smallest distance linking all lakes in a minimum spanning tree; this is a standard method in PCNM analysis. MEM variables were created from a patristic distance matrix (Cain and Harrison, 1960) along the river network, all edges having equal lengths of 1. In the same spirit, AEM variables were constructed with all edges having equal weights.

The adjusted coefficient of determination (R^2_a) corrects for the number of explanatory variables in the model and for the number of observations. It provides an unbiased estimate, in RDA, of the real contributions of the independent variables to the explanation of a response data table (Peres-Neto et al., 2006). This statistic was used in Table 2 to compare the results of the five RDA models. R^2_a values are not given for CCA because canonical analysis packages (e.g., Canoco, or the 'vegan' R-language library) do not produce them yet due to its recent discovery (Peres-Neto et al., 2006) and the complexity of its calculation. The ordinary R^2 statistic was used to compare CCA results to those of the other modelling techniques, with the understanding that R^2 is biased and produces higher values when the number of explanatory variables is larger.

Results show that a larger proportion of the diet variation (R^2, R^2_a) is explained by the AEM spatial model than by any of the other models presented in Table 2. The AEM model, which is constructed from the edges of the river network, accounts for a very large portion ($R^2_a = 63.6\%$) of the variation in brook trout diet composition among the lakes. That model may have captured both geomorphological differences among portions of the river network and differences among brook trout populations, which migrated from lake to lake along the network. In 1994, Magnan et al. had mostly related the variation in trout diet to environmental variables, including morphological characteristics of the lakes, and a smaller fraction to the spatial distribution of the lakes on the map of the Mastigouche Reserve (through geographic polynomial analysis) or along the river network (through CCA based on nodes). AEM modelling presents a strong improvement over the modelling methods that were available at the time.

Fig. 6 presents a triplot of the AEM model. This model clearly shows 3 groups of lakes, with perhaps a few intermediate ones: lakes with brook trout populations dominated by zoobenthos eaters (lower right), by zooplankton eaters (lower left), and by generalists whose diet includes benthos, zooplankton, as well as prey-fish, aquatic insects, and terrestrial insects (upper central). Bourke et al. (1997) associated

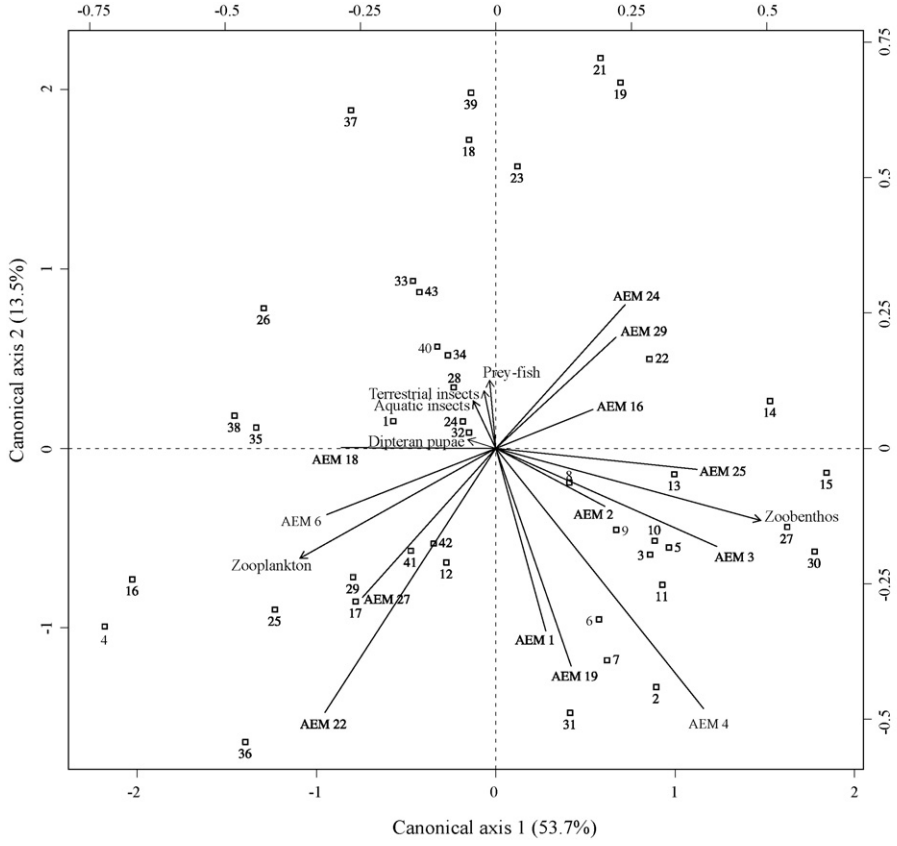


Fig. 6 – RDA triplot (axes 1 and 2) showing the 42 lakes (open squares labelled 1–43), 9 prey categories (five are shown by arrows; the other 4 were very short and contributed little to the ordination plane), and 13 AEM eigenfunctions (lines). The only significant axes were 1 and 2.

these three lake groups with three morphologically differentiable forms of brook trout, which they called the benthic, pelagic, and generalist individuals. The pelagic form is morphologically distinguishable from the benthic and generalist individuals. The RDA triplot (Fig. 6) also shows that AEM variables 16, 22, 24, 27, and 29 model the lakes dominated by the pelagic form of brook trout (zooplankton eaters) whereas AEM eigenfunctions 2, 3, 4 and 25 model lakes dominated by benthic individuals (zoobenthos eaters). AEM variables 1 and 19 are more suited to model lakes dominated by generalists, which have negative scores along these variables.

For the subset of 37 lakes, Lacasse and Magnan (1992) had shown the same differences among brook trout populations using biotic (presence of the creek chub *Semotilus atromaculatus* and the white sucker *Catostomus commersoni*, and zooplankton community structure) and abiotic variables (sampling date, morphoedaphic index, importance of rock outcrops). They emphasized the direct and indirect impacts of white suckers, explaining that their presence selectively favours the pelagic form of brook trout. This conclusion was strengthened by Bourke et al. (1999) who found that creek chubs have the same impact on the distribution of brook trout forms, although to a lesser extent. These observations support the hypothesis that polymorphism is promoted by relaxation of interspecific competition.

AEM analysis lends itself to different types of graphical representation. First, one can draw bubble-plot maps of the significant, individual AEM variables (not shown). A more parsimonious representation is obtained by plotting RDA fitted site scores on maps; the fitted site scores of canonical axes 1 and 2 are plotted as bubble maps in Fig. 7a and b. Another, more concise representation is obtained by partitioning the lakes using their RDA fitted site scores (all axes) by *K*-means (Fig. 7c). The partition was mapped for four groups. Each group of lakes is a good representation of the different forms of brook trout. Since this partition explains 63.6% (and not 100%) of the variance of the brook trout diet composition, the three groups of trout are not perfectly recognizable on that map.

A note has to be added regarding the way the selection of spatial variables was done for this illustration. Contrary to the method proposed in Blanchet et al. (in press), we used the whole set of AEM eigenfunctions in the forward selection procedure. We decided to proceed in that way because we were expecting both positive and negative autocorrelation to be of importance in this example. The finest scale of the sampling being a lake, two lakes that were geographically close could be very different with regard to the dietary habits of brook trout. The same theoretical consideration would also apply to MEM eigenfunctions.

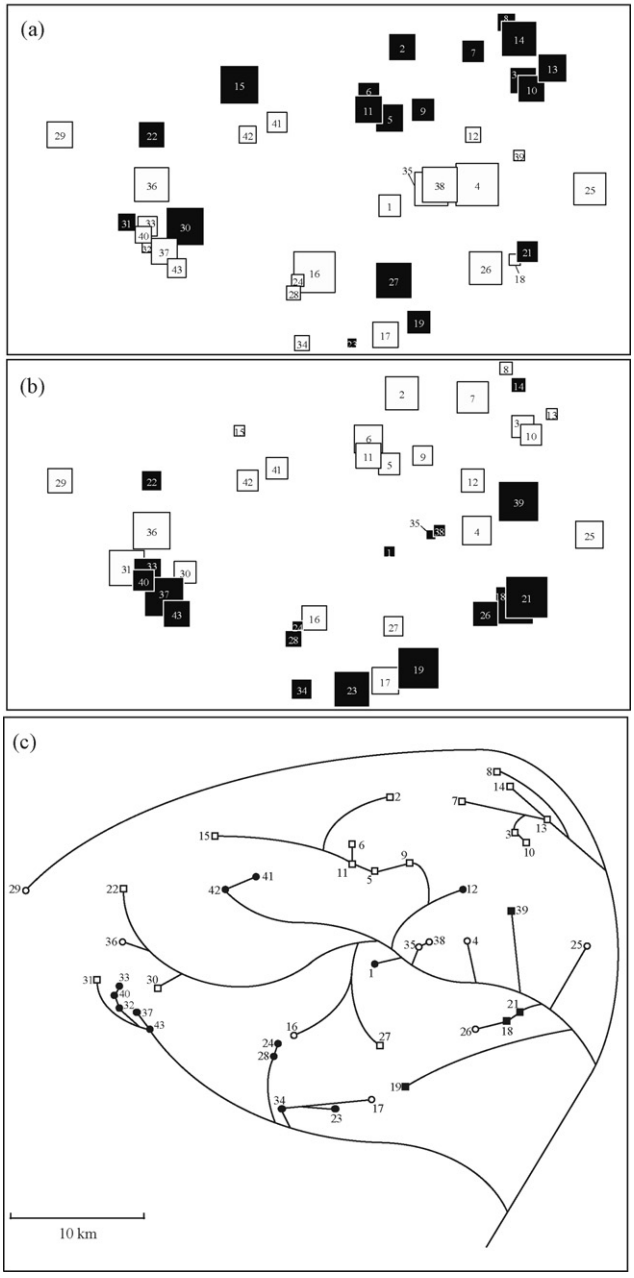


Fig. 7 – Bubble plot maps of the RDA fitted site scores for (a) axis 1 and (b) axis 2; black square bubbles are positive, white bubbles are negative; square size is proportional to the absolute values represented. (c) Four-group *K*-means partition of the lakes plotted on the river network map using symbols.

5. Discussion

The objective of spatial modelling using geographic eigenfunctions differs from that of standard canonical modelling using only environmental variables as the explanatory table. Magnan et al. (1994) did both types of modelling, acknowledging the fact that the presence of spatial structures in communities is of great interest: it indicates that some process has been at work to create these structures. Ecologists now understand that spatial structures can be produced by two different mechanisms (Legendre and Legendre, 1998, p. 11; Fortin and Dale, 2005, pp. 214–216): they may be the result of spatial dependence induced by environmental forcing variables onto the community under study (niche-based processes); they may also be the result of the dynamics of the community itself (neutral processes). These two types of generating processes can often be distinguished because they act at different spatial scales. Variation partitioning, mentioned in the

first paragraph of Section 1, further allows ecologists to determine how much of the community variation explained by the environmental variables is also spatially structured.

The AEM framework allows researchers to construct with great flexibility spatial variables (eigenfunctions) corresponding to hypothesized asymmetric generating processes. Three types of information are needed to create AEM eigenfunctions. (1) The geographic coordinates of the sites under study. (2) A connexion diagram linking the sites together. How to obtain that information may be obvious when one considers a river network, as in our ecological example. It may also be less clearly defined, especially when finer-scale phenomena are investigated. We suggest using prior information, if at hand, to construct the connexion diagram. Current velocity, water depth, presence of water masses, geological and historical events, etc. could be of great interest to construct an asymmetric connexion network well suited for a particular data set. (3) Last and most important, a direction in which the asymmetrical process operates. With these three types of information, a binary sites-by-edges table (**E**) can be constructed. This table, with or without weights added to the edges, can be directly used to construct AEM eigenfunctions.

The AEM framework is not limited to model systems influenced by a single directional process. There are indeed situations where two opposite directional processes may be at work. In a river or ocean current system for example, larvae may come down with the current whereas predators may be coming up. The resulting distribution of larvae may result from the combined action of these two processes. We could, for example, construct a first set of AEM eigenfunctions corresponding to the downstream process and a second set of AEM eigenfunctions corresponding to the upstream process. These two AEM tables could then be used in variation partitioning to estimate their relative contributions to the explanation of the variation of the response data.

The combination of connexion diagrams and weighted edges offers a broad range of possibilities to create AEM eigenfunctions for a particular set of site coordinates. Ecological theory or knowledge about the context of the study (i.e., the way processes are spreading among the sampling units) should prevail when choosing a particular set of weights. Note, however, that this approach does not *a priori* warrant that the results obtained from the AEM analysis are the best possible in terms of variance explained, but the tradeoff is favourable if the AEM have been built to represent prior knowledge as closely as possible.

Variables known to be influenced by directional processes have been studied using various approaches. For example, custom-designed models may be developed after an extensive study of a system; Abril and Abdel-Aal (2000) used this approach to model pollutant dispersion in the Suez Canal. This approach gives good results; however, the model built cannot easily be generalized. Geostatistics have also been used extensively to model asymmetric spatial processes. Using variograms in more than one direction to model anisotropic processes is very common in that discipline (Isaaks and Srivastava, 1989; Cressie, 1993; Wackernagel, 2003). However, this approach produces asymmetric but non-directional models and it cannot be applied to multivariate situations. Recent research in geostatistics has focused on the development of spatial statistics specialized to model stream networks (Ver Hoef et al., 2006; Peterson et al., 2007). These can also only be applied to univariate situations. Also, the models are built specifically for stream networks, whereas the AEM framework can provide asymmetric spatial modelling variables for any situation where there is evidence for the presence of an asymmetric process influencing the spatial distribution of the species under study.

In the last few years, numerous methodological developments have been proposed to model space more accurately. Up to very recently, the trend in spatial modelling was to develop and use methods that could model space for any ecological situation. Trend surface, PCNM and MEM analyses are good examples of those general methods. Presently, researchers are developing new techniques that are specialized for modelling the effects of particular generating processes. The AEM method follows that trend. As was mentioned earlier, when no directional process is involved, there is no point in constructing spatial variables through the AEM framework. The core of this article is to show that AEM variables are more efficient than MEM variables when a directional spatial process is considered.

The particularities of AEM eigenfunctions make it possible for this framework to be used in other fields of research. One future direction would be to use this method to address phylogenetic research questions since it is well suited to model tree-like structures, with and without reticulations.

6. Supplements

An R package called “AEM” is available online. It contains all the functions used to perform the analyses presented in this paper.

Acknowledgements

We are grateful to Prof. P. Magnan, Université du Québec à Trois-Rivières, who allowed us to use the brook trout diet data for illustration of the AEM method. This research was supported by NSERC grant no. OGP0007738 to P. Legendre.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ecolmodel.2008.04.001.

REFERENCES

Abril, J.M., Abdel-Aal, M.M., 2000. A modelling study on hydrodynamics and pollutant dispersion in the Suez Canal. Ecological Modelling 128, 1–17.

Anderson, M.J., Legendre, P., 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.

Barthélemy, J.-P., Guénoche, A., 1988. Les arbres et les représentations des proximités. Masson, Paris.

Berge, C., 1958. Théorie des graphes et ses applications. Dunod, Paris.
Blanchet, F.G., Legendre, P., Borcard, D., in press. Forward selection of explanatory variables. Ecology.
Borcard, D., Legendre, P., 2002. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. Ecological Modelling 153, 51–68.
Borcard, D., Legendre, P., Avois-Jacquet, C., Tuomisto, H., 2004. Dissecting the spatial structure of ecological data at multiple scales. Ecology 85, 1826–1832.
Borcard, D., Legendre, P., Drapeau, P., 1992. Partialling out the spatial component of ecological variation. Ecology 73, 1045–1055.
Bourke, P., Magnan, P., Rodriguez, M.A., 1997. Individual variations in habitat use and morphology in brook charr. Journal of Fish Biology 51, 783–794.
Bourke, P., Magnan, P., Rodriguez, M.A., 1999. Phenotypic responses of lacustrine brook charr in relation to the intensity of interspecific competition. Evolutionary Ecology 13, 19–31.
Cain, A.J., Harrison, G.A., 1960. Phyletic weighting. Proceedings of the Zoological Society of London 135, 1–31.
Cressie, N.A.C., 1993. Statistics for Spatial Data. John Wiley and Sons, New York.
Dray, S., Legendre, P., Peres-Neto, P.R., 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). Ecological Modelling 196, 483–493.
Fortin, M.-J., Dale, M.R.T., 2005. Spatial Analysis—A Guide for Ecologists. Cambridge University Press, Cambridge.
Griffith, D.A., 2000. A linear regression solution to the spatial autocorrelation problem. Journal of Geographical Systems 2, 141–156.
Griffith, D.A., Peres-Neto, P.R., 2006. Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses. Ecology 87, 2603–2613.
Huston, M.A., 1996. Biological Diversity: The Coexistence of Species on Changing Landscapes. Cambridge University Press, Cambridge.
Isaaks, E.H., Srivastava, R.M., 1989. An Introduction to Applied Geostatistics. Oxford University Press, New York.
Kludge, A.G., Farris, J.S., 1969. Quantitative phylectics and the evolution of anurans. Systematic Zoology 18, 1–32.
Lacasse, S., Magnan, P., 1992. Biotic and abiotic determinants of the diet of brook trout, *Salvelinus fontinalis*, in lakes of the Laurentian Shield. Canadian Journal of Fisheries and Aquatic Sciences 49, 1001–1009.
Legendre, P., 1990. Quantitative methods and biogeographic analysis. In: Garbary, D.J., South, R.G. (Eds.), Evolutionary Biogeography of the Marine Algae of the North Atlantic. Springer-Verlag, Berlin, pp. 9–34.
Legendre, P., Borcard, D., 2006. Quelles sont les échelles spatiales importantes dans un écosystème? In: Droesbeke, J.-J., Lejeune, M., Saporta, G. (Eds.), Analyse statistique des données spatiales. Éditions Technip, Paris, pp. 435–442.
Legendre, P., Fortin, M.-J., 1989. Spatial pattern and ecological analysis. Vegetatio 80, 107–138.
Legendre, P., Legendre, L., 1998. Numerical Ecology, 2nd English ed. Elsevier Science BV, Amsterdam.
Magnan, P., Rodriguez, M.A., Legendre, P., Lacasse, S., 1994. Dietary variation in a freshwater fish species: relative contribution of biotic interactions, abiotic factors, and spatial structure. Canadian Journal of Fisheries and Aquatic Sciences 51, 2856–2865.
Manly, B.F.J., 1997. Randomization, Bootstrap and Monte Carlo Methods in Biology, 2nd ed. Chapman and Hall, London.
Oksanen, J., Kindt, R., Legendre, P., O'Hara, R.B., 2007. vegan: community ecology package. R package version 1.9-25. URL http://cran.r-project.org/.
Peres-Neto, P.R., Legendre, P., Dray, S., Borcard, D., 2006. Variation partitioning of species data matrices: estimation and comparison of fractions. Ecology 87, 2614–2625.
Peterson, E.E., Theobald, D.M., Hoef, J.M.V., 2007. Geostatistical modelling on stream networks: developing valid covariance matrices based on hydrologic distance and stream flow. Freshwater Biology 52, 267–279.
R Development Core Team, 2007. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
Rao, C.R., 1964. The use and interpretation of principal component analysis in applied research. Sankhyaá (The Indian Journal of Statistic) Series A 26, 329–358.
Ronquist, F., 1996. Matrix representation of trees, redundancy, and weighting. Systematic Biology 45, 247–253.
Sidak, Z., 1967. Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association 62, 626–633.
ter Braak, C.J.F., 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67, 1167–1179.
Ver Hoef, J.M., Peterson, E., Theobald, D., 2006. Spatial statistical models that use flow and stream distance. Environmental and Ecological Statistics 13, 449–464.
Wackernagel, H., 2003. Multivariate Geostatistics—An Introduction with Applications, 3rd ed. Springer, New York.