

Direction des bibliothèques

AVIS

Ce document a été numérisé par la Division de la gestion des documents et des archives de l'Université de Montréal.

L'auteur a autorisé l'Université de Montréal à reproduire et diffuser, en totalité ou en partie, par quelque moyen que ce soit et sur quelque support que ce soit, et exclusivement à des fins non lucratives d'enseignement et de recherche, des copies de ce mémoire ou de cette thèse.

L'auteur et les coauteurs le cas échéant conservent la propriété du droit d'auteur et des droits moraux qui protègent ce document. Ni la thèse ou le mémoire, ni des extraits substantiels de ce document, ne doivent être imprimés ou autrement reproduits sans l'autorisation de l'auteur.

Afin de se conformer à la Loi canadienne sur la protection des renseignements personnels, quelques formulaires secondaires, coordonnées ou signatures intégrées au texte ont pu être enlevés de ce document. Bien que cela ait pu affecter la pagination, il n'y a aucun contenu manquant.

NOTICE

This document was digitized by the Records Management & Archives Division of Université de Montréal.

The author of this thesis or dissertation has granted a nonexclusive license allowing Université de Montréal to reproduce and publish the document, in part or in whole, and in any format, solely for noncommercial educational and research purposes.

The author and co-authors if applicable retain copyright ownership and moral rights in this document. Neither the whole thesis or dissertation, nor substantial extracts from it, may be printed or otherwise reproduced without the author's permission.

In compliance with the Canadian Privacy Act some supporting forms, contact information or signatures may have been removed from the document. While this may affect the document page count, it does not represent any loss of content from the document.

Université de Montréal

Analyse spatiale en écologie : développements méthodologiques

par

Guillaume Blanchet

Département de Sciences Biologiques

Faculté des Arts et des Sciences

Mémoire présenté à la Faculté des Études Supérieures

en vue de l'obtention du grade de

Maître ès science (M.Sc.)

Août 2007

© Guillaume Blanchet 2007



Identification du Jury

Université de Montréal
Faculté des Études Supérieures

Ce mémoire intitulé :

Analyse spatiale en écologie : développements méthodologiques

présenté par

Guillaume Blanchet

a été évalué par un jury composé des personnes suivantes :

François-Joseph LapointePrésident-rapporteur

Pierre LegendreDirecteur de recherche

Daniel BoisclairMembre du jury

Mémoire accepté le 26 novembre 2007

RÉSUMÉ

Ce travail, à deux volets, propose d'une part [1] l'amélioration d'une méthode de sélection de variables afin qu'elle soit mieux adaptée à des variables spatiales orthogonales et, d'autre part, [2] les cartes de vecteurs propres asymétriques qui constituent une nouvelle méthode permettant de générer des variables spatiales en considérant l'asymétrie spatiale d'un processus écologique.

[1] La méthode progressive de la régression pas à pas est souvent utilisée en écologie pour sélectionner un jeu réduit de variables explicatives. C'est une méthode efficace pour construire un modèle statistique concis. Par contre, son utilisation avec des variables spatiales construites dans le cadre des cartes de vecteurs propres de Moran (ou *Moran's eigenvector maps*, MEM) a tendance à surestimer la quantité de variances expliquée et à gonfler l'erreur de type I. Le premier chapitre de ce travail propose une innovation à cette méthode de sélection pour pallier à ces problèmes. Une procédure en deux étapes est développée. En premier lieu, un test global en utilisant tout le jeu de variables spatiales doit être réalisé. Si, et seulement si, le test global est significatif, la méthode progressive de la régression pas à pas peut être appliquée. Pour éviter la surestimation de la variance expliquée, la régression pas à pas doit être faite en utilisant deux critères d'arrêt, soit (1) le critère de réjection alpha, ce qui est commun pour tout type de régression pas à pas, et (2) le coefficient de détermination multiple ajusté (R^2_a) calculé avec toutes les variables spatiales disponibles. Lorsqu'une variable spatiale fait dépasser le seuil fixé pour l'un ou l'autre des deux critères, cette variable est rejetée et la sélection s'arrête.

[2] La répartition spatiale des espèces, tant animales que végétales, terrestres qu'aquatiques, est influencée par de nombreux facteurs, comme les gradients physiques et biogéographiques. Par exemple, la direction du vent dominant ou d'un courant induit des gradients qui peuvent influencer la répartition spatiale de nombre d'espèces alors que des événements historiques (e.g. une glaciation) peuvent créer des gradients biogéographiques. À ce jour, aucune technique de modélisation spatiale n'a été développée qui considère

l'asymétrie d'un processus contrôlant lorsqu'une étude de la répartition spatiale est faite le long d'un gradient. Le deuxième chapitre de ce travail présentera une nouvelle méthode modélisant la répartition des espèces dans l'espace en présence d'un processus asymétrique connu. Cette méthode est une extension des MEM. La méthode produit les cartes de vecteurs propres asymétriques (ou *asymmetric eigenvector maps*, AEM).

Chacun des chapitres de ce travail sera illustré par des données écologiques réelles. Le premier chapitre est illustré par l'analyse de données du Parc national Bryce Canyon (Utah, États-Unis d'Amérique) alors que le second est illustré par l'analyse de données de contenus stomacaux d'ombles de fontaine (*Salvelinus fontinalis*) provenant de 42 lacs de la réserve Mastigouche, Québec, Canada.

Mots-clés : méthode progressive de la régression pas à pas, asymétrie spatiale, coordonnées principales de matrice de voisinage (PCNM), carte de vecteurs propres de Moran (MEM), carte de vecteurs propres asymétriques (AEM), réserve Mastigouche, *Salvelinus fontinalis*, Bryce Canyon National Park.

SUMMARY

This two-chapter work presents first an improvement of the forward selection procedure that is better suited for orthogonal spatial variables. It also proposes a new method to generate spatial variables, which considers the spatial asymmetry of an ecological process. These variables are called asymmetric eigenvector maps.

The first chapter of this work proposes a new way of using forward selection that is well adapted to eigenfunction-based spatial filtering methods. The classical forward selection procedure carried out on orthogonal spatial variables presents a highly inflated rate of type I error. To prevent this, we propose a two-step procedure. First, a global test using all spatial variables must be carried out. If, and only if, the global test is significant, one can proceed with forward selection. Furthermore, to prevent overestimation of the explained variance, the forward selection has to be carried out with two stopping criteria: (1) the usual alpha level of rejection and (2) the adjusted coefficient of multiple determination (R^2_a) calculated with all spatial variables. When forward selection identifies a variable that brings one or the other criterion over the fixed threshold, this variable is rejected and the procedure stops.

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); such factors are known to influence the spatial distributions of many species. No spatial modelling technique has been developed to this day that considers the asymmetry of the controlling factors when studying species distributions along a gradient. Here will be presented a new spatial modelling method that can model species spatial distributions generated by a known asymmetric process. This method is an eigenfunction-based spatial filtering method; it pertains to the same general framework as Moran's eigenvector maps (MEM) analysis. The new method is called asymmetric eigenvector maps (AEM). To

illustrate how this new method works, AEM are compared to MEM through simulations and with an ecological example where a known asymmetric forcing is present.

An ecological illustration is presented for each chapter. The first chapter uses plant data gathered in Bryce Canyon National Park (Utah, USA). The second chapter uses dietary habits of brook trout (*Salvelinus fontinalis*) sampled in 42 lakes in the Mastigouche Reserve, Québec.

Key words: Forward selection, spatial asymmetry, principal coordinates of neighbor matrices (PCNM), Moran's eigenvector maps (MEM), asymmetric eigenvector maps (AEM), Mastigouche Reserve, *Salvelinus fontinalis*, Bryce Canyon National Park.

TABLE DES MATIÈRES

RÉSUMÉ	ii
SUMMARY	iv
TABLE DES MATIÈRES	vi
LISTE DES TABLEAUX	vii
LISTE DES FIGURES	viii
REMERCIEMENTS	xii
INTRODUCTION	1
CHAPITRE 1 : Forward selection of explanatory spatial variables	8
Abstract	10
Introduction	10
Difference between PCNM and MEM variables	11
Forward selection: a huge type I error	12
Global Test: a way to achieve a correct type I error rate	14
Structured Response Variables: towards an accurate modeling	15
Example: Bryce Canyon Data	19
Discussion	20
CHAPITRE 2 : Modelling spatial asymmetry in ecological data	33
Abstract	34
Introduction	35
Method	36
Simulation study	39
Ecological illustration	44
Discussion	48
CONCLUSION	66
BIBLIOGRAPHIE	67

LISTE DES TABLEAUX

CHAPITRE 1

Table 1: Percentage of time when all the variables used to create the response variable, and only those, were chosen by the forward selection procedure.

CHAPITRE 2

Table 1: Weighting function and α parameter giving the highest explained variance when modeling each structure of each set of simulation, with AEM or MEM. Results were obtained after 1000 simulations done with each combination of weighting function (2) and α parameter (10). The same response variables were used to compare AEM and MEM variables.

Table 2: Comparison of spatial models of brook trout diet obtained from 7 different modeling methods. While Magnan et al. (1994) used a subset of 37 lakes and a cutoff level of $\alpha = 0.10$ in their forward selection in CCA, we used the full set of 42 lacs and a cutoff level of $\alpha = 0.05$ for the results presented in this table.

LISTE DES FIGURES

INTRODUCTION

Figure 1 : Principe de l'analyse spatiale PCNM, basée sur les coordonnées principales d'une matrice de voisinage (matrice tronquée de distances euclidiennes entre les sites). (Modifié de Legendre et Borcard 2006).

Figure 2 : Principe de l'analyse spatiale MEM.

CHAPITRE 1

Figure 1 : Result of 5000 simulations of forward selection when only alpha is used as a stopping criterion. The response variable is random normal. (a) R^2_a for each simulation, black = BEM, grey = PCNM. The mean of the 5000 simulations is presented with a line going through the distribution. (b) Number of PCNMs selected by forward selection. (c) Number of BEMs selected by forward selection.

Figure 2 : Type I error of BEMs on series of 100 data points randomly selected from four distributions. For each distribution, 5000 independent simulation were completed. The error bars represent 95% confidence intervals.

Figure 3 : Variation of R^2_a when randomly selected spatial variables are added to a model already containing the correct explanatory variables. Spatial variables were added one at a time until none was left to add. 5000 simulations were done. Whiskers: extreme values. (a) Results for PCNMs. (b) Results for positively autocorrelated BEMs. (c) Results for negatively autocorrelated BEMs.

Figure 4 : Comparison of a forward selection done on PCNMs with both the R^2_a and alpha level as stopping criteria (a-b, e-f, i-j) with one where only the alpha criterion (c-d, g-h, k-l) was used. Three different situations are presented: (1) the standard deviation of the deterministic part of the response variable is the same as the standard deviation of the error (a-d), (2) the standard deviation of the error is 0.25 times that of the standard deviation of the deterministic part (e-h) and (3) the standard deviation of the error is 0.001 times that of the standard deviation of the deterministic part (i-l). The left-hand side presents the correct selections made by the forward selection, i.e., the variables selected were the ones used to create the response variable. The right-hand side shows the bad selections, i.e. the variables selected were not the ones used to create the response variable. 5000 simulations were run for each magnitude of error.

CHAPITRE 2

Figure 1: Schematic representation of AEM analysis using a fictive example. Sites are linked together with a connection diagram (a-b), which in turn will be used to construct a sites by link matrix (c). Weights can be added to the links (column) of this matrix. Descriptors (AEM variables) are then constructed through the calculation of SVD or PCA eigenvectors (d). Construction of AEM variables can also be done through the calculation of a distance matrix and the calculation of eigenvectors via PCoA.

Figure 2: Type I error of AEM analysis (b, d) with connection diagram and points (a) and (c) respectively. No weights were used in (a), whereas the inverse of the distance was used as weight in (c). The large arrows present the direction of the asymmetry considered for (a) and (c). Response variables were randomly selected for each point from four different distributions. Each run consisted of 5000 independent simulations. The error bars in (b) and (d) represent 95% confidence intervals.

Figure 3: (a) Connection diagram used to create AEM and MEM variables. Arrows represent direction of influence of each site on others; these directions were considered to construct AEM variables but not for MEM variables. (b) Eight basic structures (S1 to S8) used to generate response variables. Numbers are weights added, prior to adding random normal noise, to one whole line (1 to 10) of the regular grid used (a) for simulations. Each pair of structure presents a symmetric (even numbered structure) and an asymmetric (odd numbered structure) pattern in the generation of the data.

Figure 4: Variance explained (R^2_a) for the best set of AEM (full line) and MEM (dotted line) variables for each of the 8 structures presented in Figure 3. (a-c) present results of univariate simulations where the error parameter was randomly chosen from a normal distribution with a standard deviation of 1, 2, and 3 respectively. (d) presents results of multivariate simulations where the error parameter was randomly chosen from a normal distribution with a standard deviation selected from a uniform distribution with a minimum of 1 and a maximum of 3. Error bars represent 95% confident intervals. Each run consists of 1000 independent simulations. Lines linking error bars were plotted to prevent confusion between results of AEM and MEM analysis.

Figure 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. Adapted from Magnan et al. (1994).

Figure 6: RDA triplot (axes 1 and 2) showing the 42 lakes (open square), 9 prey categories (5 are shown by arrows, the other 4 were very short and contributed little to the ordination plane), and 13 AEM variables (lines). Axes 1 and 2 were the only significant axes.

Figure 7: Bubble plot maps of the RDA fitted site scores for (a) axis 1 and (b) axis 2; black bubbles are positive, white bubbles are negative; circle sizes are proportional to the absolute values represented. (c) Four groups *K*-means partition of the lakes plotted on the river network map.

REMERCIEMENTS

En premier lieu, je souhaite remercier mon directeur de recherche, Pierre Legendre. Il aurait été impossible pour moi de faire ce projet sans son support intellectuel et financier. Son enthousiasme contagieux pour la recherche, la bonne bouffe, le bon vin, les Macs... et tant d'autres choses n'ont fait que rendre l'expérience plus mémorable. Le dynamisme du Labo Legendre a été pour moi un environnement de croissance intellectuelle fantastique sur tous les plans (de l'histoire à la statistique bayésienne en passant par les recettes de homard au chocolat). Pierre a été pour moi plus qu'un mentor, il a aussi été un ami qui m'a ouvert les yeux sur le monde.

Très près derrière, je désire aussi remercier Daniel Borcard pour tout le support qu'il m'a accordé autant pour mes travaux de recherche que pour l'enseignement de la biostatistique. Son esprit un peu tordu pour régler des problèmes de tout ordre ainsi que les discussions élaborées sur de si nombreux sujets m'ont permis d'évoluer dans mon cheminement personnel.

Aussi, j'aimerais remercier les membres du Labo Legendre qui furent présents lors de mon séjour : Marie-Hélène Ouellette, Sébastien Durand, Pedro R. Peres-Neto, Stéphane Dray, Einar Heegaard, Elaine Hooper, Miquel DeCacères Ainsa, Philippe Casgrain et Charleyne Bachraty; vous m'avez fait voir de nombreuses facettes de la recherche scientifique en rendant le sujet abordable pour moi... qui étais fraîchement sorti du monde de la physiologie animale.

Ensuite, je voudrais rendre un hommage posthume aux quelques virus de la grippe qui ont réussi à m'affecter pendant ces deux dernières années. Malgré les périodes de faiblesse qu'ils m'ont apportées, ces derniers m'ont obligé à prendre quelques jours de repos qui ont été plus qu'appréciés.

Pour terminer, je dédie ce mémoire à mes parents Richard et Diane. Leur support moral, les moments de détente qu'ils m'ont imposés et les valeurs qu'ils m'ont transmises ont rendu l'achèvement de ce travail possible.

INTRODUCTION

L'importance de l'hétérogénéité spatiale en écologie est bien connue et ce depuis longtemps (Kolasa et Rollo 1991). Par contre, les méthodes permettant d'étudier ces phénomènes sont arrivées beaucoup plus tardivement. En 1989, Legendre et Fortin ont publié un article qui s'avéra être un point tournant pour l'analyse spatiale en écologie. Ils présentèrent plusieurs méthodes provenant de domaines extérieurs à l'écologie permettant d'expliquer l'impact des phénomènes spatiaux sur la répartition des communautés végétales. Ces méthodes ont été utilisées dans d'autres sphères de l'écologie comme la limnologie (e.g. Cooper et al. 1997), l'océanographie (e.g. Planque et al. 1997), l'écologie animale (e.g. Bergin 1992).

Ensuite, plusieurs écologistes plus versés dans la statistique et les mathématiques se sont lancés dans le développement de méthodes pour analyser spécifiquement l'espace en écologie. Un bon exemple de développement méthodologique fait par un écologiste pour mieux comprendre les phénomènes spatiaux en écologie est la partition de la variation (Borcard et al. 1992). Cette méthode a été développée originalement pour mieux comprendre quelle portion de la variance expliquée est uniquement due aux variables spatiales d'un modèle, uniquement aux variables environnementales, ainsi qu'à une combinaison de ces deux groupes de variables.

Avec la partition de la variation, il devenait impératif de trouver une façon de générer des variables permettant de bien modéliser la répartition spatiale des organismes. La méthode la plus simple permettant de générer ce genre de variables, connue à l'époque du développement de la partition de la variation, était de calculer un polynôme de deuxième ou de troisième ordre à partir des coordonnées géographiques (Legendre 1990). Ceci consiste à prendre les coordonnées XY des sites et les élever au premier (X et Y), au deuxième (X, Y, XY, X² et Y²) ou au troisième degré (X, Y, XY, X², Y², X²Y, XY², X³ et Y³). Ce polynôme formait le tableau des variables explicatives dans une régression ou une analyse canonique. Il devenait donc possible d'analyser des structures spatiales dans un contexte écologique. Ce type de méthode a par contre un défaut : pour pouvoir modéliser la

distribution d'organismes à une échelle relativement fine, il est nécessaire de disposer d'un polynôme extrêmement long. Quoique mathématiquement possible, l'utilisation d'un polynôme d'ordre supérieur à trois présente plusieurs problèmes. La robustesse d'un test statistique peut être diminuée si trop de variables sont incorporées dans un modèle. Ce problème est particulièrement important lorsque le nombre d'observations est faible, ce qui est fréquemment le cas en écologie. Un autre problème lié à l'utilisation d'un polynôme de grand ordre (plus que trois) est la difficulté qu'on peut avoir à interpréter l'effet de ces variables sur le tableau-réponse.

Pour pallier aux inconvénients qu'engendre l'utilisation des polynômes des coordonnées géographiques, Borcard et Legendre (2002) ont développé les coordonnées principales de matrices de voisinage (*Principale coordinate of neighbour matrices, PCNM*, en anglais); l'acronyme anglais sera utilisé dans le reste du texte pour éviter toute confusion avec l'acronyme français des cartes de vecteurs de Moran. Les PCNM sont des variables orthogonales issues d'une décomposition spectrale d'une matrice de distances tronquée calculée à partir des coordonnées géographiques des sites d'échantillonnage (Figure 1). Une matrice de distance tronquée consiste en une matrice de distance où toutes les distances plus grandes que la plus grande distance dans la chaîne permettant de relier tous les sites ensemble sont remplacées par une valeur très grande (4 fois la plus grande distance considérée, ou plus). Elles ont l'avantage de permettre de déceler des variations à échelle fine et ce même si un nombre très restreint de sites ont été échantillonnés. Elles peuvent être utilisées dans des contextes très variés. Borcard et al. (2004) présentent plusieurs situations écologiques très différentes où l'analyse PCNM a produit des résultats très intéressants.

Il a ensuite été montré par Dray et al. (2006) que les PCNMs font partie d'un cadre général, les cartes de vecteurs propres de Moran (*Moran's eigenvector maps, MEM* en anglais); l'acronyme anglais sera utilisé dans le reste du texte pour éviter toute confusion avec celui des coordonnées principales de matrices de voisinage. Alors que les PCNM sont uniquement basées sur les distances entre les sites échantillonnés, le cadre des MEM

présente une façon de créer des variables où non seulement les distances entre les sites peuvent être prises en considération, mais aussi le nombre de voisins; les sites peuvent être reliés entre eux par un diagramme de connexions permettant de définir quels sites ont une influence les uns sur les autres. La figure 2 présente schématiquement la construction de variables spatiales construites dans le cadre des MEM. Les MEMs permettent une très grande flexibilité qu'aucune autre méthode d'analyse spatiale n'avait jusqu'alors.

Les PCNMs, comme les MEMs, ont aussi leurs défauts et leurs limites. Ces deux méthodes permettent de générer un nombre très important de variables spatiales. Pour les PCNMs, il est fréquent de voir $2n/3$ variables générées, n étant le nombre de sites échantillonnés. Pour les MEMs, il arrive souvent qu'il y ait $(n - 1)$ variables générées, ce qui est encore pire, puisque avec autant de variables, un test statistique est impossible à faire par manque de degrés de liberté.

Ces deux méthodes se veulent généralistes : elles peuvent être utilisées dans toutes les situations où l'on souhaite modéliser la structure spatiale des données échantillonnées. Malheureusement, certaines situations requièrent des méthodes plus spécifiques. Les PCNMs et les MEMs tentent de modéliser la répartition spatiale d'organismes sans prendre en considération des connaissances qu'on pourrait posséder *a priori* sur un milieu étudié. Par exemple, si on tente de modéliser la répartition spatiale d'organismes dans une rivière à l'aide des PCNMs ou des MEMs, même si ces dernières sont très flexibles, aucune de ces méthodes ne permet de prendre en considération le fait qu'un courant puisse influencer de façon directionnelle la répartition spatiale des organismes étudiés.

Dans toutes ces méthodes de modélisation spatiale, un grand nombre de fonctions sont générées pour décrire les relations spatiales entre les sites d'échantillonnage. Il est intéressant dans certains cas de réduire le nombre de variables spatiales explicatives des données écologiques à l'aide d'une des méthodes de sélection menant à un modèle parcimonieux. Un modèle parcimonieux a plus de pouvoir prédictif (Gauch 1993, 2003). Cela est désirable par exemple lors de la formulation de sous-modèles spatiaux correspondant à des échelles spatiales différentes ou encore lorsqu'on veut représenter les

variables spatiales dans un diagramme d'ordination. La méthode couramment employée en analyse canonique est la sélection ascendante (*forward selection*, en anglais) des variables explicatives. Or on sait que cette méthode est trop libérale; en d'autres termes, elle a tendance à incorporer dans le modèle des variables qui n'ont qu'un effet aléatoire au niveau de la population statistique. Parce que nous analysons un échantillon de taille réduite, ces variables peuvent, par hasard, modéliser une partie du bruit qui se trouve dans les données.

Les deux chapitres de ce travail ont pour but de résoudre les deux problèmes mentionnés ci-dessus.

Le premier chapitre propose une nouvelle méthode de sélection de variables spatiales orthogonales. Il a pour but d'avertir les utilisateurs de cette méthode à propos des comportements capricieux de la sélection progressive lorsque cette dernière est utilisée pour sélectionner des variables spatiales orthogonales. Ce chapitre propose aussi une nouvelle procédure de sélection progressive pour sélectionner des variables provenant du cadre des MEM où le nombre de variables spatiales explicatives est $(n - 1)$.

Cette nouvelle procédure sera validée à l'aide de simulations. Un jeu de données sur la biodiversité des plantes vasculaires du Parc national de Bryce Canyon (Utah, États-Unis d'Amérique) sera utilisé pour illustrer comment cette nouvelle procédure réagit dans une situation écologique réelle.

Le second chapitre de ce travail présente une nouvelle méthode pour générer des variables spatiales. Il est bien connu que la répartition spatiale des espèces peut être influencée par un ou des gradients des variables environnementales (Huston 1996). Beaucoup de gradients sont induits par des processus spatiaux asymétriques. Malgré les développements méthodologiques importants qui ont permis de mieux comprendre comment les structures spatiales influencent la répartition des espèces, aucune méthode ne considère les processus asymétriques. La méthode développée ici entre dans le cadre des méthodes de filtrage spatial basées sur le calcul de valeurs et de vecteurs propres, concept développé par Griffith et Peres-Neto (2006).

À échelle fine comme large, la répartition spatiale des espèces est souvent structurée selon un ou plusieurs gradients, biotiques et/ou abiotiques. Nous proposons d'utiliser des variables spatiales qui sont asymétriques par construction pour étudier la répartition spatiale de communautés d'espèces qui sont influencées par des gradients. Dray et al. (2006) déplorent l'absence de méthode considérant l'asymétrie spatiale; notre article servira à combler cette lacune dans la littérature. Comme pour les MEMs, les variables asymétriques présentées dans ce chapitre proviennent d'un cadre général très flexible permettant de générer des variables spatiales asymétriques. Les variables créées dans ce cadre s'appellent des cartes de vecteurs propres asymétriques (*asymmetric eigenvector maps*, AEM, en anglais); l'acronyme anglais sera utilisé ici. Ces variables se veulent appropriées pour des situations où les processus environnementaux influençant les organismes étudiés possèdent une asymétrie spatiale connue (e.g. dans une rivière, un fleuve ou un courant marin). Ce nouveau développement sera validé par des simulations créées dans un contexte bidimensionnel. Un jeu de données sur les contenus stomacaux des ombles de fontaine (*Salvelinus fontinalis*) dans 42 lacs de la réserve Mastigouche sera utilisé pour illustrer l'utilisation des AEMs dans une étude écologique réelle. Une comparaison entre les AEMs et plusieurs autres méthodes, dont les MEMs et les PCNMs, sera faite pour ce même jeu de données.

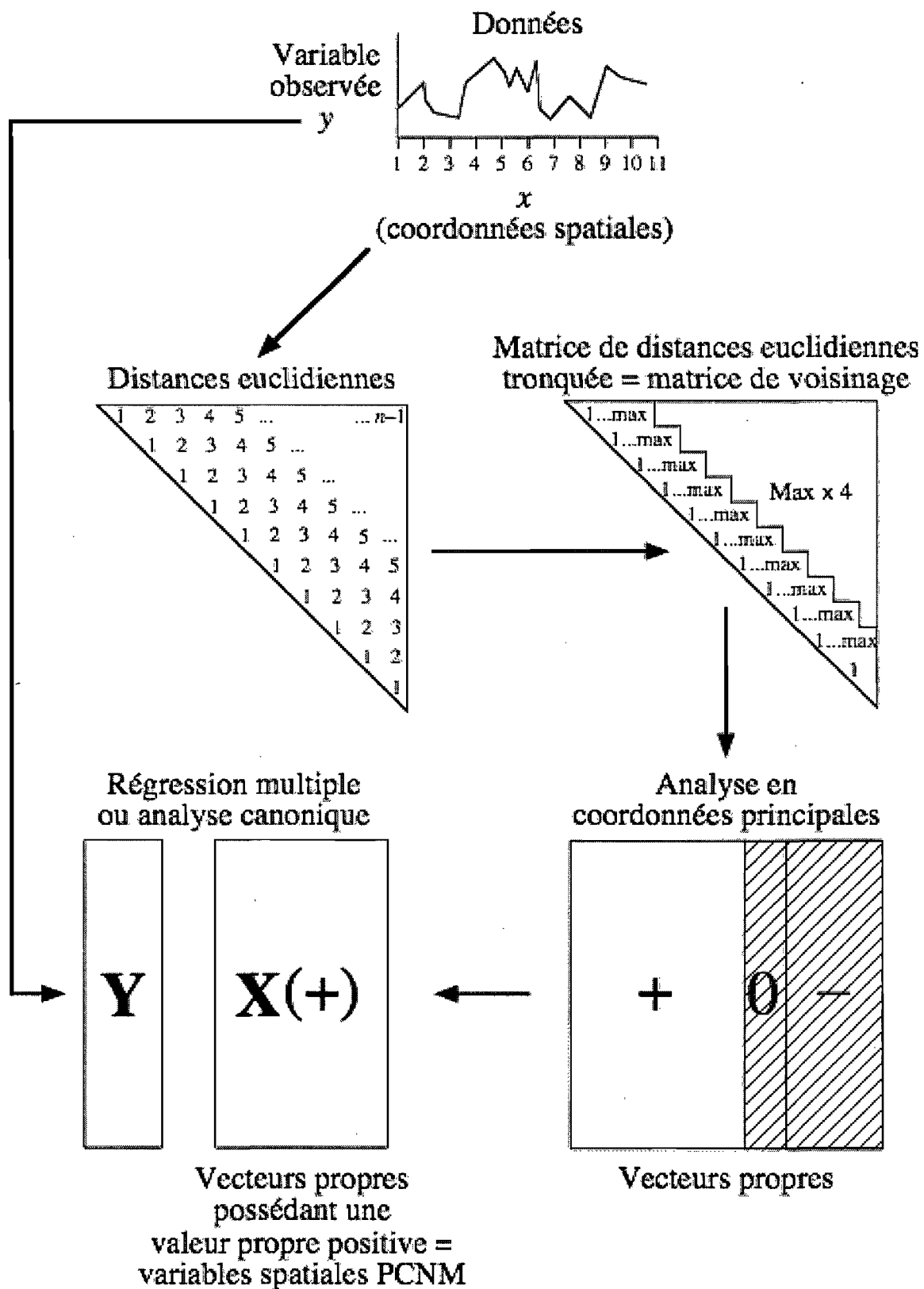


Figure 1


Chapitre 1

Forward selection of explanatory spatial variables

Forward selection of explanatory spatial variables

F. GUILLAUME BLANCHET^{1,2}, PIERRE LEGENDRE¹, AND DANIEL BORCARD¹

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

² Corresponding address: F. Guillaume Blanchet, Département de sciences biologiques,
Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Québec, Canada H3C
3J7. E-mail: 

Abstract. This report proposes a new way of using forward selection that is well adapted to eigenbased spatial filtering methods. The classical forward selection carried out on orthogonal spatial variables presents a very inflated type I error. To prevent this, we propose a two steps procedure. First, a global test using all spatial variables must be carried out. If, and only if, the global test is significant, one can proceed with a forward selection. Furthermore, to prevent overestimation of the explained variance, the forward selection has to be carried out with two stopping criteria (1) the usual alpha level of rejection and (2) the adjusted coefficient of multiple determination (R^2_a) calculated with all spatial variables. When forward selection identifies a variable that brings one or the other criterion over the fixed threshold this variable is rejected and the procedure is stopped. This new technique is validated with simulations and an ecological example is presented with data from Bryce Canyon National Park (Utah, USA).

Key words: Principal coordinates of neighbor matrices (PCNM), Moran's eigenvector maps (MEM), spatial analysis, simulations, type I error

INTRODUCTION

Since the introduction of principal coordinates of neighbor matrices (PCNM) (Borcard and Legendre 2002, Borcard et al. 2004) and of Moran's eigenvector maps (MEM) (Dray et al. 2006), ecologists have been faced with the problem of having to handle large numbers of spatial explanatory variables in their analyses. In their concluding remarks, Bellier et al. (2007) stated: "PCNM requires methods to choose objectively the composition, number, and form of spatial submodels". We propose a new method for selecting spatial submodels for those types of variables. The new method is completely independent of the user's knowledge of the data under study.

An automatic selection procedure is used in most cases to select a subset of explanatory variables objectively. Having fewer variables that explain almost the same amount of variance is interesting; it retains enough degrees of freedom for testing the F -

statistic in situations where the number of observations is small because observations are very costly. Furthermore, a parsimonious model has greater predictive power (Gauch 1993, 2003). One method very often used for selecting variables in ecology is forward selection. It presents the great advantage of working even when the initial dataset has more explanatory variables than sites, which is often the case in ecology. Since forward selection is being used more and more to select spatial variables (e.g. Borcard et al. 2004, Brind'Amour et al. 2005, Duque et al. 2005, Telford and Birks 2005, Halpern and Cottenie 2007), it is this report's goal to warn researchers against the sometimes capricious behavior of forward selection when selecting orthogonal spatial variables. We also propose a new forward selection procedure to select variables constructed through an eigenfunction-based spatial filtering method where the number of spatial explanatory variables is equal to $(n - 1)$, where n is the number of objects.

The procedure will be presented and validated with the help of simulated data. To illustrate how it reacts on real ecological data, we shall use the Bryce Canyon National Park (Utah, USA) dataset.

DIFFERENCE BETWEEN PCNM AND MEM VARIABLES

MEMs are a general framework to construct the many variants of orthogonal, eigenvector-based spatial variables like PCNMs and distance-based eigenvector maps (Dray et al. 2006). For example, PCNMs are constructed on the basis of a distance criterion. This is not necessarily the case of other MEMs that can be constructed based on a connection diagram, a number of neighbors, etc. Detailed explanation of the construction of PCNMs and MEMs are presented in Borcard and Legendre (2002) and Dray et al. (2006) respectively.

In this report, we will use two types of spatial variables out of the MEM framework to present our new approach of forward selection and investigate its properties by numerical simulations. The first type is PCNMs because they are the most widely used at the moment

in ecology (e.g. Duque et al. 2005, Kohler et al. 2006). PCNM is an eigen-based spatial decomposition method that creates spatial variables (PCNM eigenfunctions) through a truncated distance matrix initially constructed from the geographical coordinates of the study sites. The other type is the simplest construction from the MEM framework, which we call binary eigenvector maps (BEM) in this report. BEM are constructed from a connexion diagram, which, in the particular case of a transect, links all sites from left to right. No weights will be added to the links in the simulations presented in this paper. The connexion matrix derived from the connexion diagram is used directly to build spatial variables through a principal coordinate analysis (PCoA). All simulations and analyses were carried out on an irregular transect of 100 sites. For irregularly spaced sites, PCNMs and BEMs represent two extreme types in the MEM framework (Dray et al. 2006).

FORWARD SELECTION: A HUGE TYPE I ERROR

The simulations presented below show that, when used in the traditional manner (i.e., step-by-step introduction of explanatory variables with a test of the partial contribution of each variable to be entered), forward selection of orthogonal spatial variables presents two problems: (1) an inflated type I error, and (2) an overestimation of the amount of variance explained. In a first set of simulations to measure the type I error rate, we created a random normal response variable along a transect containing 100 irregularly spaced simulated sampling sites. The site positions along the transect were created using a random uniform generator. The same transect was used for all simulations. The simulations differ in the data generated at those specific sites. PCNMs were computed from the spatial coordinates of the points along the transect, and a forward selection was carried out to identify the PCNM variables best suited to model the response variable, with a stopping α level of 0.05. To increase computation speed, we ran all analyses using a parametric forward selection procedure, adequate here because the simulated data were random normal. Parametric tests should not, however, be used with non-normal data such as tables of

species abundances. In such cases, randomization procedures should be used (Pitman, 1937a, 1937b and 1938). We repeated this procedure with 5000 independent sets of random normal data. The same simulations were repeated with BEMs.

The simulation results are presented in Fig. 1. On PCNMs the procedure behaved correctly roughly 6% of the time only, selecting no PCNM to model a random variable, i.e., the overall type I error rate was about 94%. This is astonishingly high when compared to the expected rate of 5%. Very often in the simulations (about 73% of the cases), one to four PCNMs were selected to model random noise. Sometimes, up to 14 PCNMs were admitted into the model. These results show that forward selection yields a hugely inflated type I error. When forward selection was applied to BEMs, results were even more alarming. Only once in 5000 tries did the forward selection lead to the correct result of not selecting any BEM. Almost 60% of the time, 7 to 17 BEM variables were selected incorrectly. As was the case for PCNM variables, very large numbers of BEMs were sometimes selected (up to 62). These results show that one cannot run a forward selection without some form of preliminary, overall test. They prompted us to find new criteria to improve the type I error of forward selection. This meant (1) to devise a rule to decide when it is appropriate to run a forward selection, and (2) to strengthen the stopping criterion of the forward selection to prevent it from being overly liberal.

Using numerical simulations, Ohtani (2000) has shown that the Ezekiel (1930) adjusted coefficient of multiple determination (R^2_a) is an unbiased estimator of the real contribution of a set of explanatory variables to the explanation of a response variable. Had the simulations presented above given accurate results, the adjusted coefficient of multiple determination would have been zero or close to zero all the time. In our results, after 5000 simulations, the mean of the R^2_a statistics is 13.2 % for PCNM and 47.2 % for BEM. Why do the R^2_a values diverge so strongly from zero? The fundamental problem lies with the forward selection procedure, which is exacerbated by the nature of these spatial variables. PCNM and BEM variables are structured in such a way that they are more suited than other types of variables to fit noise in the response data. The number of PCNM variables is at

most $2n/3$ whereas the number of BEM variables is $(n - 1)$. Besides being numerous, these variables are also orthogonal to one another, which means that each variable can model entirely different aspects of a response variable. Fig. 1b shows the number of PCNM variables selected during the 5000 simulations above, and Fig. 1c shows corresponding results for BEM variables. These graphs show that more BEM variables than PCNMs are incorrectly selected, simply because they are more numerous. Thioulouse et al. (1995) suggest that eigenvectors associated to small positive or negative eigenvalues are only weakly spatially autocorrelated. With that in mind, we could expect the variance in our unstructured response variables to be "explained" mainly by PCNM and BEM variables with small eigenvalues. This was not the case: results show that all eigenvectors were selected in roughly the same proportions (see Appendix A for details).

GLOBAL TEST: A WAY TO ACHIEVE A CORRECT TYPE I ERROR RATE

To prevent the inflation of type I error (our first goal), a global test needs to be done prior to forward selection. This is the first important message of this report. A global test means that all orthogonal variables created in the PCNM or BEM procedure are used together to model the response variable. However, with BEMs, there are often $n - 1$ spatial variables created. In this case no global test can be done since there are no degrees of freedom left. This problem can easily be resolved. Thioulouse et al. (1995) have argued that eigenvectors associated with high positive eigenvalues have high positive autocorrelation and describe global structures; whereas eigenvector associated with high negative eigenvalues have high negative autocorrelation and thus describe local structures. If the response variable(s) is known to be positively autocorrelated, only eigenvectors associated to positive eigenvalues should be used in the global test. On the other hand, if the response variable(s) is known to be negatively autocorrelated, only eigenvectors associated to negative eigenvalues should be used in the global test. In the case where there is no prior knowledge or hypothesis about the spatial structure of the response variable(s), two global

tests are done: one with the eigenvectors associated to negative eigenvalues and one with the eigenvectors associated to positive eigenvalues. Since two tests are done, a correction needs to be applied to the alpha level of rejection of H_0 to make sure that the test has an appropriate experimentwise rejection rate. Two corrections can be applied when there are two tests ($k = 2$), the corrections of Sidak (Sidak 1967) where $p_S = 1 - (1 - p)^k$ and Bonferroni (Bonferroni 1935) where $p_B = k \cdot p$, where p is the p-value. The Sidak correction was used in this report. Throughout this report we used a 5% rejection level.

The global test on PCNMs, as presented in the previous paragraph, has already been shown to have a correct type I error (Borcard and Legendre 2002). However, this has not been done for BEMs, so we ran simulations. Following Thioulouse et al. (1995) and after examination of the 99 BEMs obtained for $n = 100$ points, we divided the set into two subsets of roughly equal size, the 50 first BEMs (i.e. those with positive eigenvalues) being positively autocorrelated and the 49 last, negatively. Four distributions were used to construct response variables to assess the type I error. Data was randomly drawn from a normal, uniform, exponential, and exponential cubed distribution, following Manly (1997) and Anderson and Legendre (1999). A permutation test was done. We repeated the procedure 5000 times for each distribution. Results are shown in Fig. 2. In a nutshell, the rate of type I error is correct for BEMs when using a global test based on the premises presented above.

STRUCTURED RESPONSE VARIABLES: TOWARDS AN ACCURATE MODELING

When there is structure in the response variable(s), which is most often the case with real ecological data, and if, *and only if*, the global test presented above is significant, what should be done next? That depends on why the data are analyzed. If only the significance of the model and the proportion of variance explained are needed, then the procedure stops with the global test and the unbiased R^2_a of the model containing all spatial variables.

On the other hand, if the spatial structures modeled by PCNM or BEM variables need to be investigated in more detail, a selection of the important spatial variables needs to be carried out. This is where the R^2_a will be useful. As a precaution, we first checked that R^2_a is a stable statistic in the presence of additional, non-significant PCNM variables added in random order to the true explanatory variables. The following simulations were carried out. We generated PCNMs on an irregular transect containing 100 sites. To create a spatially structured response variable, five of these PCNMs were randomly selected, each of them was weighted by a number drawn from a uniform distribution (minimum = 0.5, maximum = 1), and these weighted PCNMs were added to create the deterministic component of the response variable. Finally, we added an error term drawn from a normal distribution with zero mean and a standard deviation equal to the standard deviation of the deterministic part of the response variable, to introduce a large amount of noise in the data. Multiple regressions were then calculated on the simulated response variable, first with the five explanatory PCNMs used to create the response variable (the expected value of R^2_a is then 0.5), then by adding, one at a time and in random order, each of the remaining PCNMs. This procedure was repeated 5000 times. The same procedure was run for the two sets of BEM defined above. Results are presented in Fig. 3. These results show that even when a model contains a high number of explanatory variables that are of little or no importance, the R^2_a is not affected. The reason why R^2_a were affected by forward selection in the first set of simulations presented in this report, as was shown in Fig. 1a, is that forward selection chooses the variable that is best suited to model the response regardless of the overall significance of the complete model (hence the necessity of the global test), whereas in the present simulations the model already contained the relevant explanatory variables and the next variables to enter the model were randomly selected and added no real contribution to the explanation.

In real cases, however, one does not know in advance what explanatory variables are relevant. Therefore, given that a global test is significant and a global R^2_a has been estimated, our second goal is now to prevent the selection from being overly liberal.

Preliminary simulations (not shown here, but see the Bryce Canyon example below) showed that, rather frequently, a forward selection run on a globally significant model yielded a submodel whose R^2_a was *higher* than the R^2_a of the global model. Obviously, this does not make sense.

Therefore, the second message of this paper is the following: the forward selection should be carried out with *two* stopping criteria: (1) the pre-selected significance level alpha and (2) the R^2_a statistic of the global model.

We ran a new set of simulations to assess the improvement brought by this second point. We created response variables using the same procedure as in the previous run (weighted sum of 5 randomly chosen PCNM or BEM variables), but three variants were produced, differing by the magnitude of the error term added. The first set had an error term equal to the standard deviation of the deterministic part of the response variable (as in the previous simulations), the second set had an error with standard deviation 25% that of the deterministic portion, and the last set of simulations had a negligible error term (0.001 times the standard deviation of the determinist portion). Each of these response variables was submitted to the procedure above, i.e., a global test followed, if significant, by a forward selection of explanatory variables (either PCNMs or one of the two sets of BEMs), using the double stopping criterion. Each result was compared to a result obtained when only alpha was used as the stopping criterion (as usually done). Variables selected by the forward selection were compared to the variables chosen to create the response variable. This was intended to show how efficiently forward selection can identify the correct spatial variables.

Results are presented in Fig. 4, Appendix B and Appendix C. Since PCNMs and both sets of BEMs react in the same way, Fig. 4 will be used in the discussion of all sets of spatial variables.

When error equals the standard deviation, a forward selection done with the two stopping criteria (R^2_a and alpha, Fig. 4a) rarely selected none or only one of the variables used to create the response variables (less than 1.5% of the time). Roughly 7.5% of the time 2 variables used to create the response variables were selected. This percentage exceeded

20% for three variables and 30% for four variables. In 37% of the cases, all variables used to create the response were found in the forward selection. On the other hand, the positive influence of the double stopping criterion is obvious when looking at Fig. 4b: in more than 60% of the cases no additional PCNM variable was (incorrectly) selected.

When only the alpha criterion is considered as a stopping criterion, forward selection identifies the correct variables very often (Fig. 4c). Under 1% of the time only, three variables or less that were used to create the response variable were chosen by the forward selection. However, this apparently better efficiency is counterbalanced by a much higher number of cases of bad selections: in more than 90% of the cases one or several additional variables are incorrectly selected (Fig. 4d).

The performance of forward selection improves when less error is added to the response variable (Fig. 4e to 4l), which was to be expected. Two points ought to be noticed. (1) Even when there is practically no error in the created response variables, roughly half the time, forward selection with two stopping criteria misses one of the true variables (Fig. 4i). When only the alpha criterion is used, forward selection invariably select all the good variables, even when a noticeable amount of error (25% standard deviation) is present in the data (Fig. 4k). (2) However, forward selection done with only the alpha criterion selects wrong variables, often more than one, in about 90% of the cases even when response variables are almost error free (Fig. 4l).

It is also interesting to see how many times, in each procedure, all the variables used to create the response variable, and only those, were chosen by the forward selection (Table 1). Again, results are very similar for PCNMs and positively and negatively autocorrelated BEMs; they will thus be discussed together. When half of the variation in the response variable is random noise (error term = standard deviation), the “perfect” selection is achieved roughly 10% of the time when R^2_a and alpha are used together. This result drops to less than 0.5% when only the alpha criterion is used. As expected, these results get better with less noisy response variables. However, using two stopping criteria is always better than using only one. The use of only the alpha criterion results in slightly more than 7% of

"perfect" selections when almost no noise is present in the response variable. The score is 17% when both the R^2_a and the alpha criteria are used. This better performance is due to the success of the double stopping criteria in preventing "wrong" variables to enter the model.

EXAMPLE: BRYCE CANYON DATA

To show how this new way to run forward selection behaves in a real multivariate situation, we used data from Bryce Canyon National Park (Utah, USA) (Roberts et al. 1988). The response table is composed of 169 vascular plants species sampled at 159 sites. 83 PCNMs variables were created on the basis of the site coordinates. The truncation distance was 2573.4 universal transverse mercator units (UTM). The global test was done on the linearly detrended response variables with 999 permutations and was significant (p-value < 0.001). The R^2_a calculated with all PCNMs was 26.4%. When a forward selection (999 permutations) was done with only the alpha criterion as stopping rule, 24 PCNMs were selected before the procedure stopped. However, the R^2_a calculated with those 24 PCNMs was 31.5%, i.e., a value higher than the R^2_a of the complete model. When R^2_a was added into the selection procedure as an additional stopping criterion, the number of PCNMs selected dropped to 14 (with an R^2_a of 26.4%). Therefore, based on the simulations presented above, it can be supposed that the addition of a second stopping rule prevented several unwanted PCNM variables to be admitted into the model. Furthermore, since the last of the 14 variables to enter the model explained about 0.6% variance, the procedure did not prevent any important variable to be included. It is not the purpose of this report to discuss this example in more detail, but we are confident that the more parsimonious model resulting from our improved selection procedure would be less noisy and therefore would be easier to interpret (Gauch 2003).

DISCUSSION

Carrying out a global test including all spatial variables available is not only important, it is necessary to obtain an overall correct type I error. We showed that the particular global test devised when there are too many spatial variables present, as was the case for BEMs, produces a correct type I error. But is the variance explained by a global model influenced by the obviously too numerous variables when orthogonal spatial variables are used? In other words: does the R^2_a properly correct for these particular types of spatial variables? Even though Fig. 3 shows that variations of explanation occur when variables are added to a model already well fitted, these variations are usually of low magnitude. Adding unimportant variables to an already well-fitted model has practically no impact on the explained variance measured by R^2_a . Thus, the use of R^2_a as an additional stopping criterion is a good choice in a forward selection procedure.

The use of our double stopping rule (R^2_a plus alpha level) has a number of impacts on the final selection. The most important one is that in all cases, fewer useless variables are selected. The selection is more realistic. However a comment raised by Neter et al. (1996, chapter 8) explains that the use of automatic selection procedures may lead to the selection of a set of variables that is not the best but which is very suitable for the response variable under study. Our new approach does not prevent such outcomes; it prevents the possibility of overexplaining a response variable by a set of "too-well-chosen" explanatory variables. The use of R^2_a in addition to the alpha criterion for the stopping procedure was shown, however, to select the best model more often.

Neter et al. (1996, chapter 8) proposed other parameters that could be used as stopping criteria: the total mean square error and the prediction sum of square. We decided to use the R^2_a because it offers the advantage of being also a measure of the explained amount of variance. Also, this parameter is well known in ecology, which is not the case for the other two proposed by Neter et al. (1996).

All the simulations in this report were carried out with only one response variable. This was done for simplicity. The new procedure of forward selection can also be used, without any modification, with multivariate response data sets, as illustrated here by the Bryce Canyon example.

The conclusions reached in this study are based on simulations. We tried to make the simulations as general as possible, even though we did not simulate all possible types of ecological data. This is always the case in simulation studies (Milligan 1996). A quick look at Hurlbert's unicorns (Hurlbert 1990) is a good example of how peculiar ecological data can be.

ACKNOWLEDGMENTS

This research was supported by NSERC grant no. 7738 to P. Legendre.

LITERATURE CITED

- Anderson, M. J., and P. Legendre. 1999. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *Journal of Statistical Computation and Simulation* **62**:271-303.
- Bellier, E., P. Monestiez, J.-P. Durbec, and J.-N. Candau. 2007. Identifying spatial relationships at multiple scales: principal coordinates of neighbour matrices (PCNM) and geostatistical approaches. *Ecography* **30**: 385-399.
- Bonferroni, C. E. 1935. Il calcolo delle assicurazioni su gruppi di teste. Pages 13-60. *Studi in Onore del Professore Salvatore Ortu Carboni*, Rome.
- Borcard, D., and P. Legendre. 2002. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling* **153**:51-68.
- Borcard, D., P. Legendre, C. Avois-Jacquet, and H. Tuosimoto. 2004. Dissecting the spatial structure of ecological data at multiple scales. *Ecology* **85**:1826-1832.

- Brind'Amour, A., D. Boisclair, P. Legendre, and D. Borcard. 2005. Multiscale spatial distribution of a littoral fish community in relation to environmental variables. *Limnology and Oceanography* **50**:465-479.
- Dray, S., P. Legendre, and P. R. Peres-Neto. 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecological Modelling* **196**:483-493.
- Duque, A. J., J. F. Duivenvoorden, J. Cavelier, M. Sanchez, C. Polania, and A. Leon. 2005. Ferns and Melastomataceae as indicators of vascular plant composition in rain forests of Colombian Amazonia. *Plant Ecology* **178**:1-13.
- Ezekiel, M. 1930. *Method of correlation analysis*. John Wiley and Sons, Inc., New York.
- Gauch, H. G. 1993. Prediction, Parsimony and Noise. *American Scientist* **81**:468-478.
- Gauch, H. G. 2003. *Scientific Method in Practice*. Cambridge University press, New York.
- Halpern, B. S., and K. Cottenie. 2007. Little evidence for climate effects on local-scale structure and dynamics of California kelp forest communities. *Global Change Biology* **13**:236-251.
- Hurlbert, S. H. 1990. Spatial-Distribution of the Montane Unicorn. *Oikos* **58**:257-271.
- Kohler, F., F. Gillet, S. Reust, H. H. Wagner, F. Gadallah, J.-M. Gobat, and A. Buttler. 2006. Spatial and seasonal patterns of cattle habitat use in a mountain wooded pasture. *Landscape Ecology* **21**:281-295.
- Milligan, G. W. 1996. Clustering validation: Results and implications for applied analyses. Pages 341-375 in P. Arabie, L. J. Hubert and G. De Soet, editors. Clustering and classification. World Scientific Publ. Co., River Edge, New Jersey.
- Manly, B. F. J. 1997. Randomization, bootstrap and monte carlo methods in biology. Second edition. Chapman and Hall, London.
- Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. 1996. Applied linear statistical models. Fourth edition. Irwin, Chicago.
- Ohtani, K. 2000. Bootstrapping R^2 and adjusted R^2 in regression analysis. *Economic Modelling* **17**:473-483.

- Pitman E. J. G. 1937a. Significance tests which may be applied to samples from any populations. Supplement to the journal of the royal statistical society. **4**: 119-130.
- Pitman E. J. G. 1937b. Significance tests which may be applied to samples from any populations. II. The correlation coefficient test. Supplement to the journal of the royal statistical society. **4**: 225-232.
- Pitman E. J. G. 1938. Significance tests which may be applied to samples from any populations. III. The analysis of variance test. *Biometrika*. **29**: 322-335.
- Roberts, D. W., W. D., and H. G. P. 1988. Plant community distribution and dynamics in Bryce Canyon National Park. United States Department of Interior National Park Service.
- Sidak, Z. 1967. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association* **62**:626-633.
- Telford, R. J., and H. J. B. Birks. 2005. The secret assumption of transfer functions: problems with spatial autocorrelation in evaluating model performance. *Quaternary Science Reviews* **24**:2173-2179.
- Thioulouse, J., D. Chessel, and S. Champely. 1995. Multivariate analysis of spatial patterns: a unified approach to local and global structures. *Environmental and Ecological Statistics* **2**:1-14.

Table 1: Percentage of time when all the variables used to create the response variable, and only those, were chosen by the forward selection procedure.

Error		PCNM	Positive BEM	Negative BEM
Standard deviation	Alpha & R^2_a	10.6 %	10.5 %	10.5 %
	Alpha	0.5 %	0.4 %	0.5 %
Standard deviation/4	Alpha & R^2_a	17 %	18.4 %	17.7 %
	Alpha	8.3 %	6.7 %	7 %
Standard deviation/1000	Alpha & R^2_a	17 %	16.8 %	17.2 %
	Alpha	8 %	6.9 %	7 %

FIGURE CAPTIONS

Fig. 1. Result of 5000 simulations of forward selection when only alpha is used as a stopping criterion. The response variable is random normal. (a) R^2_α for each simulation, black = BEM, grey = PCNM. The mean of the 5000 simulations is presented with a line going through the distribution. (b) Number of PCNMs selected by forward selection. (c) Number of BEMs selected by forward selection.

Fig. 2. Type I error of BEMs on series of 100 data points randomly selected from four distributions. For each distribution, 5000 independent simulation were completed. The error bars represent 95% confidence intervals.

Fig. 3. Variation of R^2_α when randomly selected spatial variables are added to a model already containing the correct explanatory variables. Spatial variables were added one at a time until none was left to add. 5000 simulations were done. Whiskers: extreme values. (a) Results for PCNMs. (b) Results for positively autocorrelated BEMs. (c) Results for negatively autocorrelated BEMs.

Fig. 4. Comparison of a forward selection done on PCNMs with both the R^2_α and alpha level as stopping criteria (a-b, e-f, i-j) with one where only the alpha criterion (c-d, g-h, k-l) was used. Three different situations are presented: (1) the standard deviation of the deterministic part of the response variable is the same as the standard deviation of the error (a-d), (2) the standard deviation of the error is 0.25 times that of the standard deviation of the deterministic part (e-h) and (3) the standard deviation of the error is 0.001 times that of the standard deviation of the deterministic part (i-l). The left-hand side presents the correct selections made by the forward selection, i.e., the variables selected were the ones used to create the response variable. The right-hand side shows the bad selections, i.e. the variables selected were not the ones used to create the response variable. 5000 simulations were run for each magnitude of error.

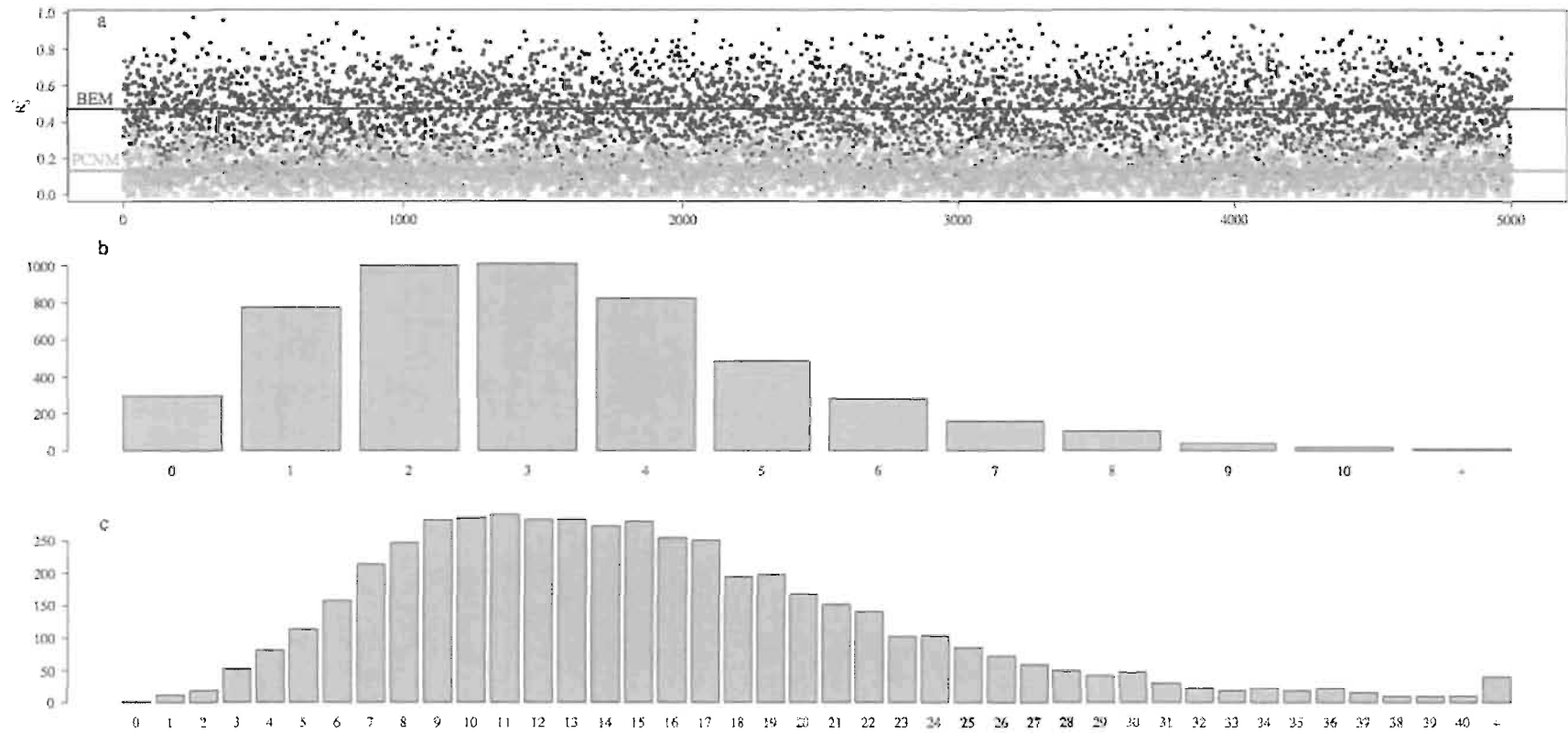


Figure 1: Result of 5000 simulations of forward selection when only alpha is used as a stopping criterion. The response variable is random normal. (a) R^2 for each simulation, black = BEM, grey = PCNM. The mean of the 5000 simulations is presented with a line going through the distribution. (b) Number of PCNMs selected by forward selection. (c) Number of BEMs selected by forward selection.

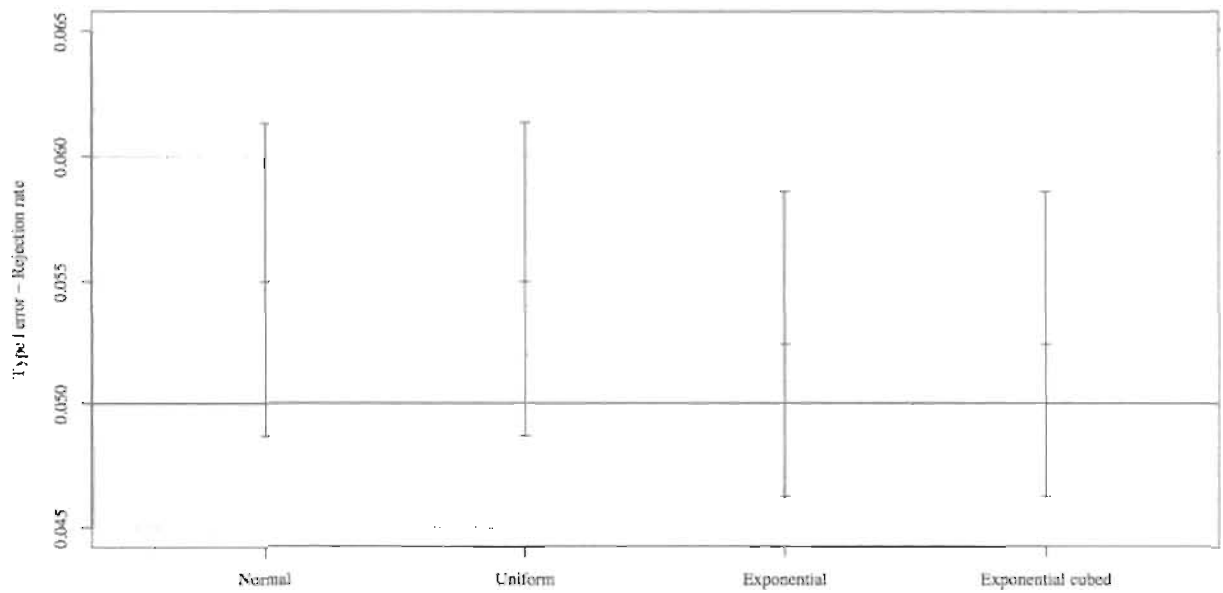


Figure 2: Type I error of BEMs on series of 100 data points randomly selected from four distributions. For each distribution, 5000 independent simulation were completed. The error bars represent 95% confidence intervals.

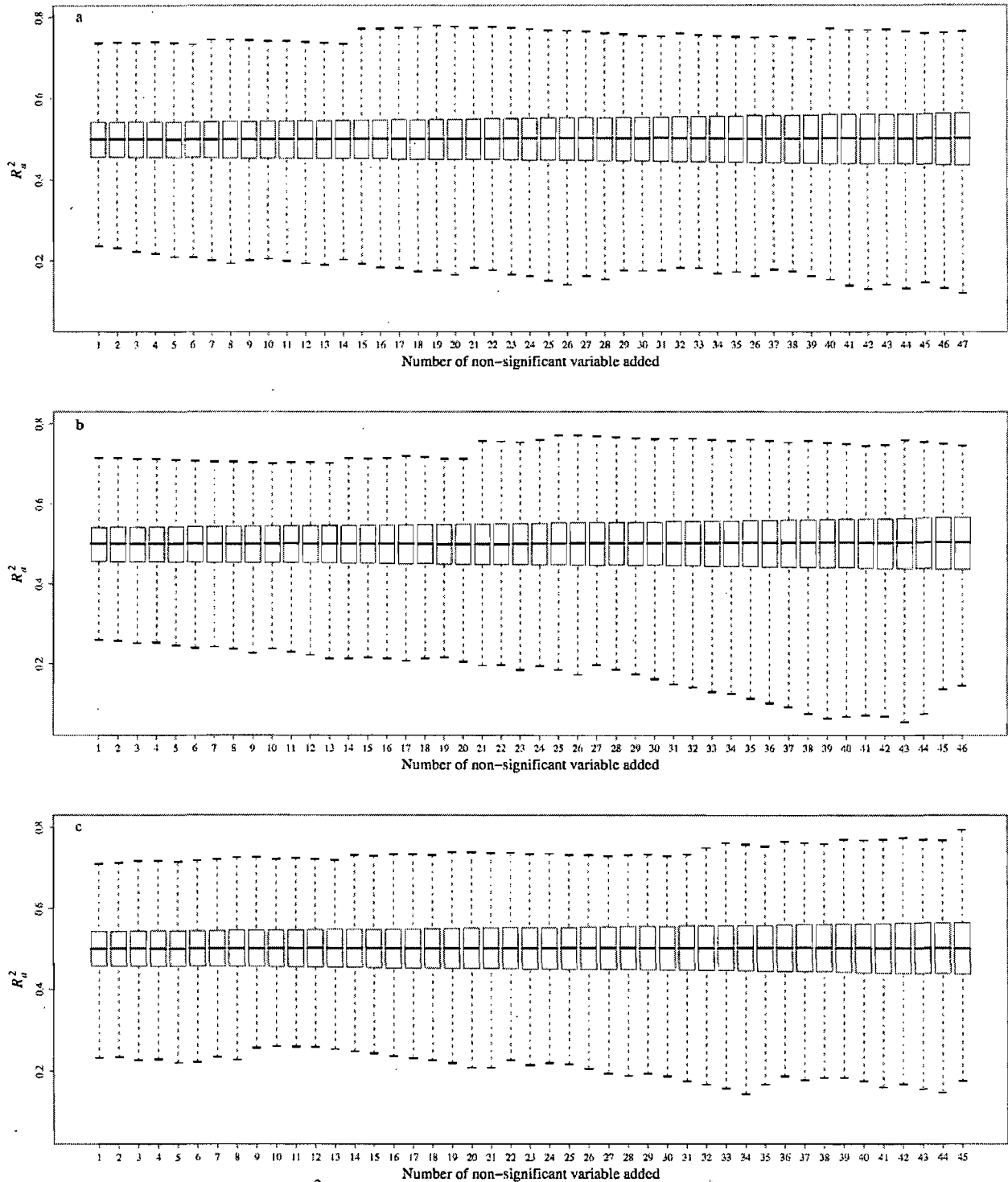


Figure 3: Variation of R_a^2 when randomly selected spatial variables are added to a model

already containing the correct explanatory variables. Spatial variables were added one at a time until none was left to add. 5000 simulations were done.

Whiskers: extreme values. (a) Results for PCNMs. (b) Results for positively autocorrelated BEMs. (c) Results for negatively autocorrelated BEMs.

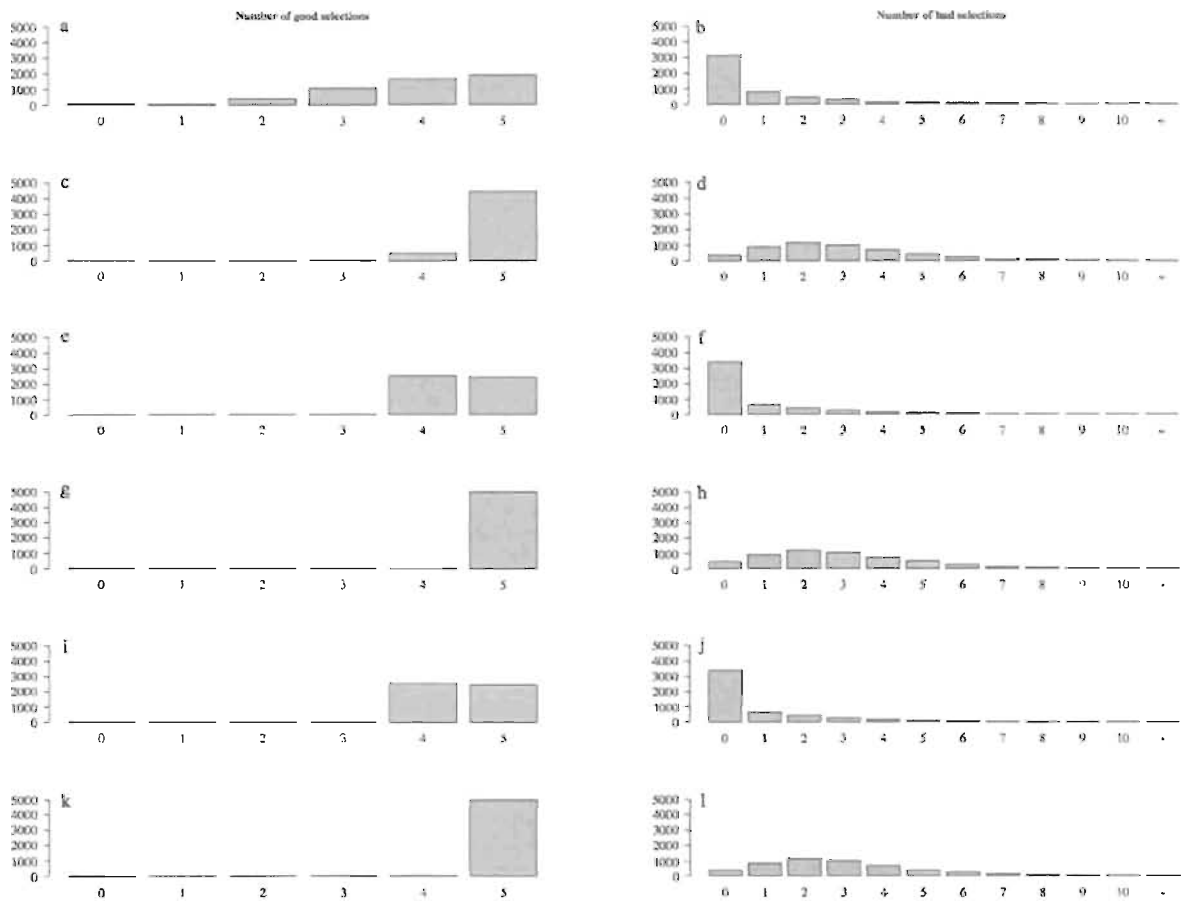
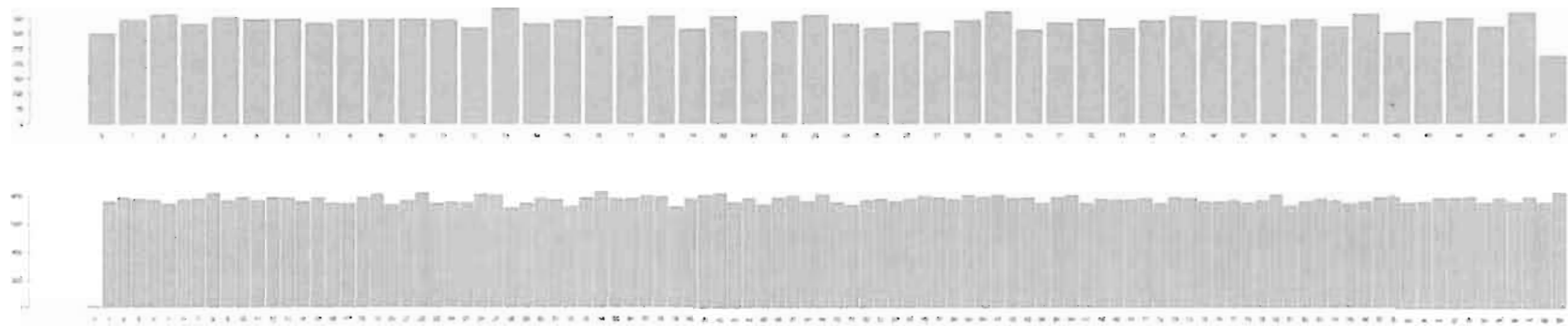
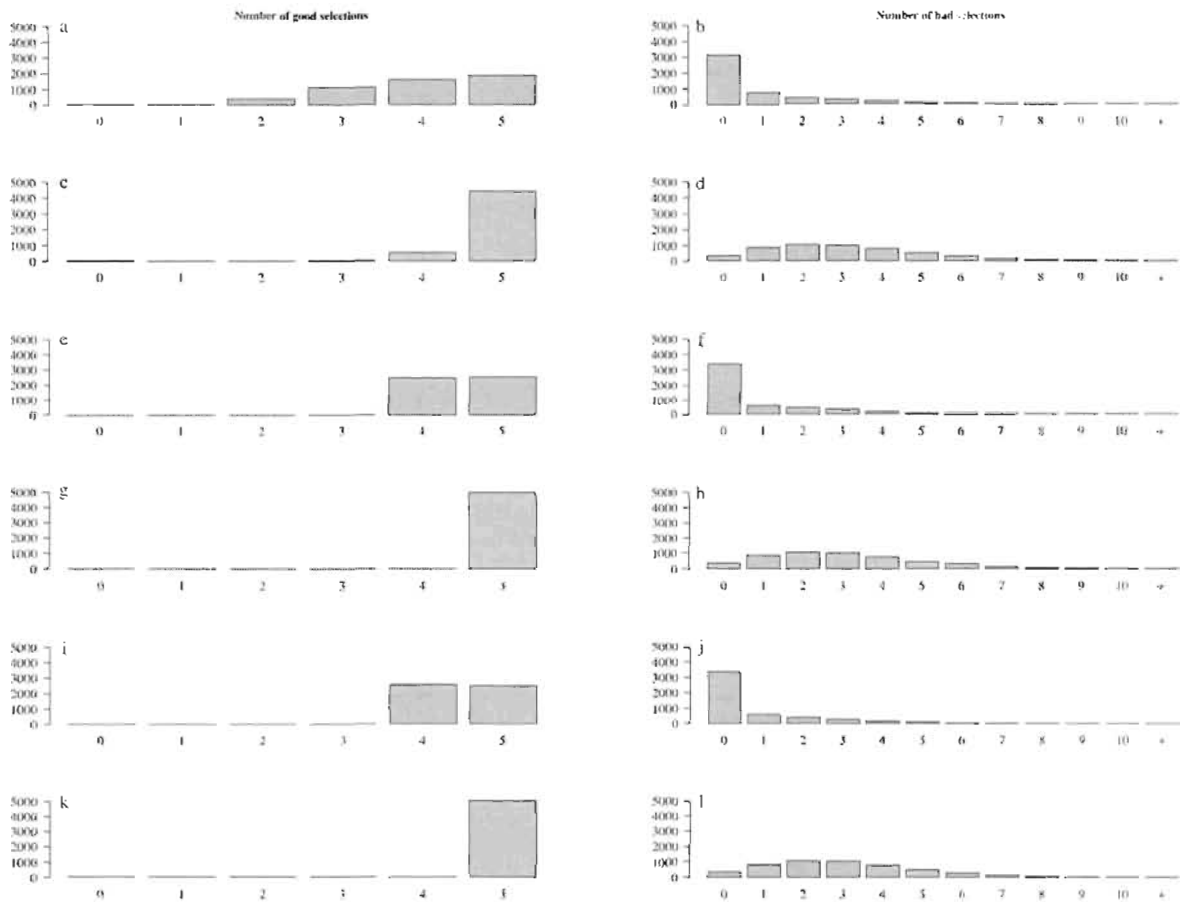


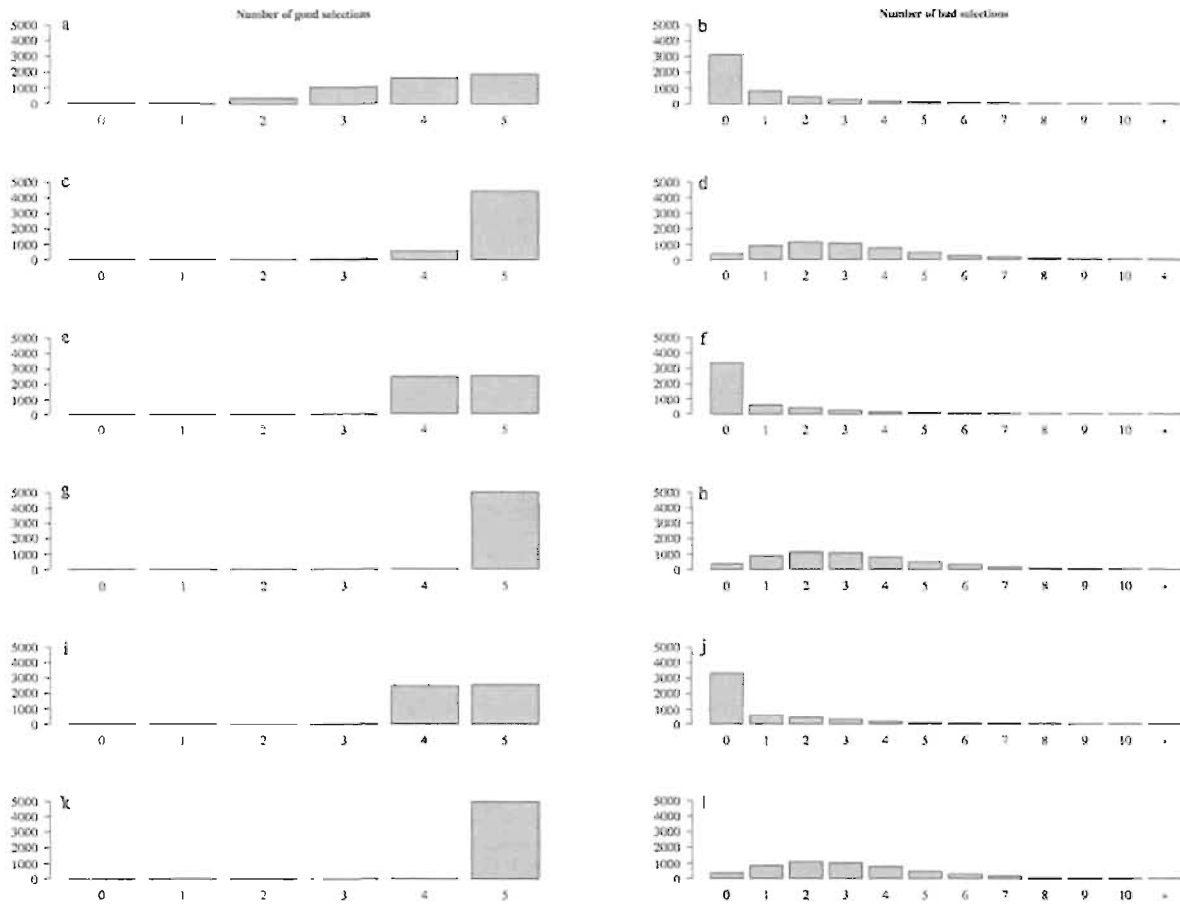
Figure 4: Comparison of a forward selection done on PCNMs with both the R^2_α and alpha level as stopping criteria (a-b, e-f, i-j) with one where only the alpha criterion (c-d, g-h, k-l) was used. Three different situations are presented: (1) the standard deviation of the deterministic part of the response variable is the same as the standard deviation of the error (a-d), (2) the standard deviation of the error is 0.25 times that of the standard deviation of the deterministic part (e-h) and (3) the standard deviation of the error is 0.001 times that of the standard deviation of the deterministic part (i-l). The left-hand side presents the correct selections made by the forward selection, i.e., the variables selected were the ones used to create the response variable. The right-hand side shows the bad selections, i.e. the variables selected were not the ones used to create the response variable. 5000 simulations were run for each magnitude of error.



Appendix A: Number of selections of each spatial variable after 5000 simulations to check for type I error of the “classical” forward selection. (a) Results for PCNMs. (b) Results for BEMs



Appendix B: Comparison of a forward selection done on positively autocorrelated BEM with both the R^2_a and alpha level as stopping criteria (a-b, e-f, i-j) with one where only the alpha criterion (c-d, g-h, k-l) was used. Three different situations are presented: (1) the standard deviation of the deterministic part of the response variable is the same as the standard deviation of the error (a-d), (2) the standard deviation of the error is 0.25 times that of the standard deviation of the deterministic part (e-h) and (3) the standard deviation of the error is 0.001 times that of the standard deviation of the deterministic part (i-l). The left-hand side presents the correct selections made by the forward selection, i.e., the variables selected were the ones used to create the response variable. The right-hand side shows the bad selections, i.e. the variables selected were not the ones used to create the response variable. 5000 simulations were run for each magnitude of error.



Appendix C: Comparison of a forward selection done on negatively autocorrelated BEM with both the R^2_a and alpha level as stopping criteria (a-b, e-f, i-j) with one where only the alpha criterion (c-d, g-h, k-l) was used. Three different situations are presented: (1) the standard deviation of the deterministic part of the response variable is the same as the standard deviation of the error (a-d), (2) the standard deviation of the error is 0.25 times that of the standard deviation of the deterministic part (e-h) and (3) the standard deviation of the error is 0.001 times that of the standard deviation of the deterministic part (i-l). The left-hand side presents the correct selections made by the forward selection, i.e., the variables selected were the ones used to create the response variable. The right-hand side shows the bad selections, i.e. the variables selected were not the ones used to create the response variable. 5000 simulations were run for each magnitude of error.

Chapitre 2

Modelling directional spatial processes in ecological data

Modelling directional spatial processes in ecological data

F. Guillaume Blanchet^{a*}, Pierre Legendre^a and Daniel Borcard^a

^aDépartement de sciences biologiques, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Québec, Canada H3C 3J7

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 asymmetry of controlling factors when studying species distributions along a gradient. 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). To illustrate how this new method works, AEM is compared to MEM analysis through simulations and 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.

* Corresponding author at: Département de sciences biologiques, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Québec, Canada H3C 3J7.

E-mail addresses: [REDACTED]

Keywords: Directional spatial process; geographic eigenfunctions; Mastigouche Reserve; Moran's eigenvector maps (MEM); *Salvelinus fontinalis*; spatial analysis; spatial autocorrelation; spatial model.

1. Introduction

It is well known that spatial distributions of species are influenced by environmental gradients (Hutson, 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 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 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 asymmetric processes by construction. Dray et al. (2006) deplored the absence of methods capable of modelling spatial asymmetry; 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 (e.g. the effects of a river network, or of currents in a sea, river, stream, or fluvial lake, on species distributions). To test the functioning and limits of the new AEM method, simulations will be carried out in a two-dimensional spatial context, where the generating process is unidirectional.

2. Method

The Dray et al. (2006) MEM method consists in the diagonalization of a spatial weighting matrix (\mathbf{W}). Matrix \mathbf{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 connections, and a weighting matrix which gives the weight associated to each pair of sites (e.g. the geographic distance or a function of the geographic distance). 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. Alternatively, compute a singular value decomposition (SVD) of the column-centred table **E**, called E_c . Decompose E_c by SVD into $U D V'$; **U** and **V** are column-orthonormal matrices and 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 connection 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. These lakes have been the subject of research for almost 20 years. Our goal with this ecological illustration is double: (1) to see how much information can be gathered from a spatial model created with AEM and (2) to present different ways to illustrate the results when AEM are used. In this example, we will assume that the directional process spatially structuring the brook trout gut contents follows the river network.

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, directly or indirectly. The sites-by-edges table \mathbf{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 \mathbf{E}' (Fig. 1c) (Ronquist, 1996). Weights can be given based on various types of known information, e.g. the inverse of the lengths of the edges.

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 tested for type I error and were compared to MEM eigenfunctions in the presence of asymmetric generating

processes, for different types of spatial structures, using the proportions of variance explained.

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 given 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 were 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 cannot carry out a test of significance using all eigenfunctions. Following Blanchet et al. (submitted), 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 than the expected value, were negatively autocorrelated. The two sets of eigenfunctions were tested separately for significance (permutation test, 999 random permutations) and their probabilities were combined using Sidak's (1967) method. Fig. 2b and 2d 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.

Simulations were also 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. A gradient is considered symmetric when the weights to be modelled are distributed evenly through the rows of the grid (even-numbered structures in Fig. 3b); otherwise it is considered asymmetric (odd-numbered structures in Fig. 3b). 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, ten 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.

Because both the AEM and MEM frameworks can create an infinite number of different spatial variables for a given 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. The connexion diagram used in all situations was the same to allow appropriate comparisons (Fig. 3a). 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 exponents of α were used. Also, in each framework (AEM, MEM), a series of spatial variables was constructed where uniform weights of 1 were given to all edges. Each set 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, meaning that a lot of random noise was present in the data (Fig 4c); 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 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 very different weights, and thus the interpretations can differ.

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 correlations 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.

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 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_a^2) 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_a^2 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_a^2) 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 ($R_a^2 = 63.6\%$) 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 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 *Castostomus commersoni* and the white sucker *Semotilus atromaculatus*, 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. 7 (a, 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. (submitted), 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.

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 the Introduction, 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 connection 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

connection diagram. Current velocity, water depth, presence of water masses, geological and historical events, etc. could be of great interest to construct an asymmetric connection 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 combination of connection diagrams and weighted edges offers a broad range of possibilities to create AEM eigenfunctions for a particular set of site coordinates. This is both good and bad. It gives flexibility to enhance the explained proportion of variance of a table of response variables; however, one can never construct all possible sets of weights and, thus, AEM eigenfunctions for a particular dataset. So, one can never be certain that the results obtained are the best that can be obtained in the AEM framework.

To guide users in their choice of a good connection network, we suggest to use prior knowledge of the studied area: river network, mapped water or wind currents, population migration routes, etc. This often helps in deciding how sites should be linked to one another. Assigning weights is a more difficult task. One solution is to use the inverse of the lengths or the squared lengths of the edges, or some other function. Weights can, generally, represent any measure of the easiness of transfer of matter or information along the edges, using prior knowledge such as current speed, dominant wind power and direction, etc. In the absence of prior information, equal weights are given to the edges.

The core of this article has been to show that AEM variables are better than MEM variables when a directional spatial process is considered. 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 specialised 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 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.

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.

References

- Anderson, M.J. and 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. and 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. and Borcard, D. Forward selection of explanatory spatial variables. Submitted to *Ecology*
- Borcard, D. and 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. and Tuomisto, H., 2004. Dissecting the spatial structure of ecological data at multiple scales. *Ecology*, 85:1826-1832.
- Borcard, D., Legendre, P. and Drapeau, P., 1992. Partialling out the spatial component of ecological variation. *Ecology*, 73:1045-1055.
- Bourke, P., Magnan, P. and 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. and 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. and Harrison, G.A., 1960. Phyletic weighting. *Proceedings of the Zoological Society of London*, 135:1-31.
- Dray, S., Legendre, P. and 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., and 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. and 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.
- Kluge, A.G. and Farris, J.S., 1969. Quantitative phyletics and the evolution of anurans. *Systematic Zoology*, 18:1-32.
- Lacasse, S. and 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. and Borcard, D., 2006. Quelles sont les échelles spatiales importantes dans un écosystème ? In: J.-J. Driesbeke, M. Lejeune and G. Saporta (Editor), *Analyse statistique des données spatiales*. Éditions TECHNIP, Paris, pp. 435-442.
- Legendre, P. and Fortin, M.-J., 1989. Spatial pattern and ecological analysis. *Vegetatio*, 80:170-138.
- Legendre, P. and Legendre, L., 1998. *Numerical Ecology*, 2nd English ed. Elsevier Science BV, Amsterdam.
- Magnan, P., Rodriguez, M.A., Legendre, P. and 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. and 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. and Borcard, D., 2006. Variation partitioning of species data matrices: estimation and comparison of fractions. *Ecology*, 87:2614-2625.
- Peterson, E.E., Theobald, D.M. and 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. *Sankhyāá (The Indian Journal of Statistic) Ser. 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.

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. 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.

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

Table 1 (Continued)

Response	Structure (S1 to S8)	AEM		MEM	
		Weighting	Exponent α	Weighting	Exponent α
		function		function	
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

Table 2 – Comparison of spatial models of brook trout diet in 42 lakes, obtained from 7 different modelling methods. Forward selection was carried out using a cutoff level of $\alpha = 0.05$.

Modelling methods	No. spatial variables in full set	No. selected spatial variables	R^2	R^2_{adj}
<i>Method based on lake geographic coordinates</i>				
CCA ¹ , 3 rd deg. polynomial	9	4 ¹	0.225	-----
RDA, PCNM analysis	24	3 ²	0.257	0.199
<i>Methods based on nodes of river network</i>				
CCA ¹ , nodes	25	5 ³	0.356	-----
RDA, nodes	25	4 ⁴	0.342	0.271
<i>Methods based on edges of river network</i>				
RDA, edges	65	9 ⁵	0.625	0.520
RDA, MEM analysis	41	11 ⁶	0.669	0.562
RDA, AEM analysis	41	13 ⁷	0.751	0.636

¹Selected monomials: X, Y, Y², X³.

²Selected PCNM variables computed from coordinates: 3, 4, 17.

³Selected nodes: 2, 9, 10, 12, 14. The nodes are shown in Fig. 1 of Magnan et al. (1994).

⁴Selected nodes: 10, 12, 14, 25.

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

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

⁷Selected AEM variables computed from edges: 1, 2, 3, 4, 6, 16, 18, 19, 22, 24, 25, 27, 29.

Figure captions

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 connection diagram (b), which in turn will be used to construct the sites-by-edges matrix \mathbf{E} (c). Weights can be attributed to the edges (column) of this matrix, representing the easiness of effect transmission between nodes (vector underneath the sites-by-edges matrix). (d) Descriptors (AEM variables, matrix \mathbf{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 \mathbf{X}) can also be obtained through the calculation of an Euclidean distance matrix followed by the computation of eigenvectors via principal coordinate analysis (PCoA).

Fig. 2 – Type I error of AEM analysis (b, d) for sampling points and connection 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.

Fig. 3 – (a) Connection 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 values added to all points on each line (1 to 10) of the diagram in (a), prior to adding random normal noise.

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.

- 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).
- Fig. 6 – RDA triplot (axes 1 and 2) showing the 42 lakes (open squares labelled 1 to 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.
- 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.

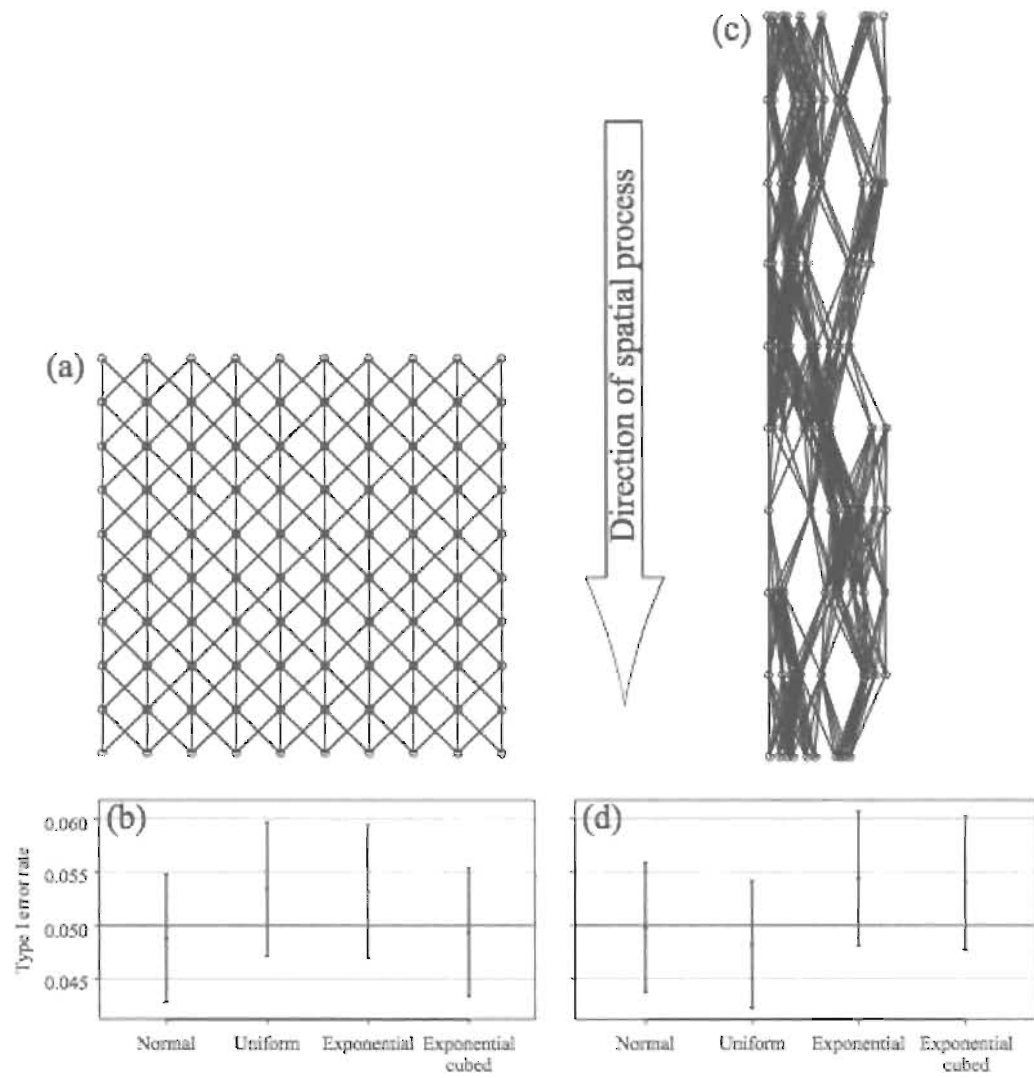


Fig. 2

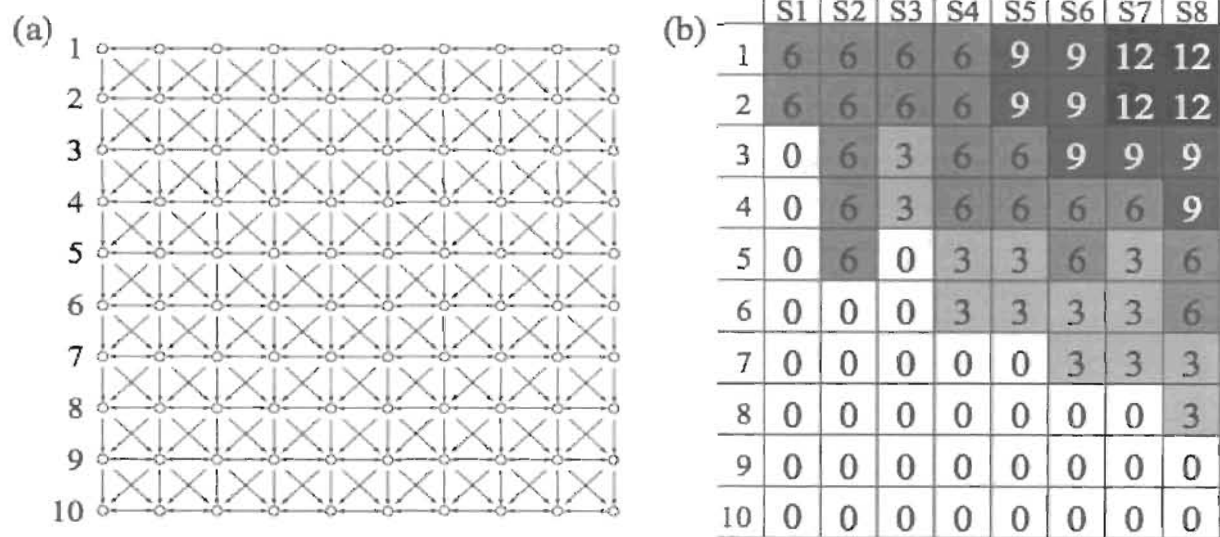


Fig. 3

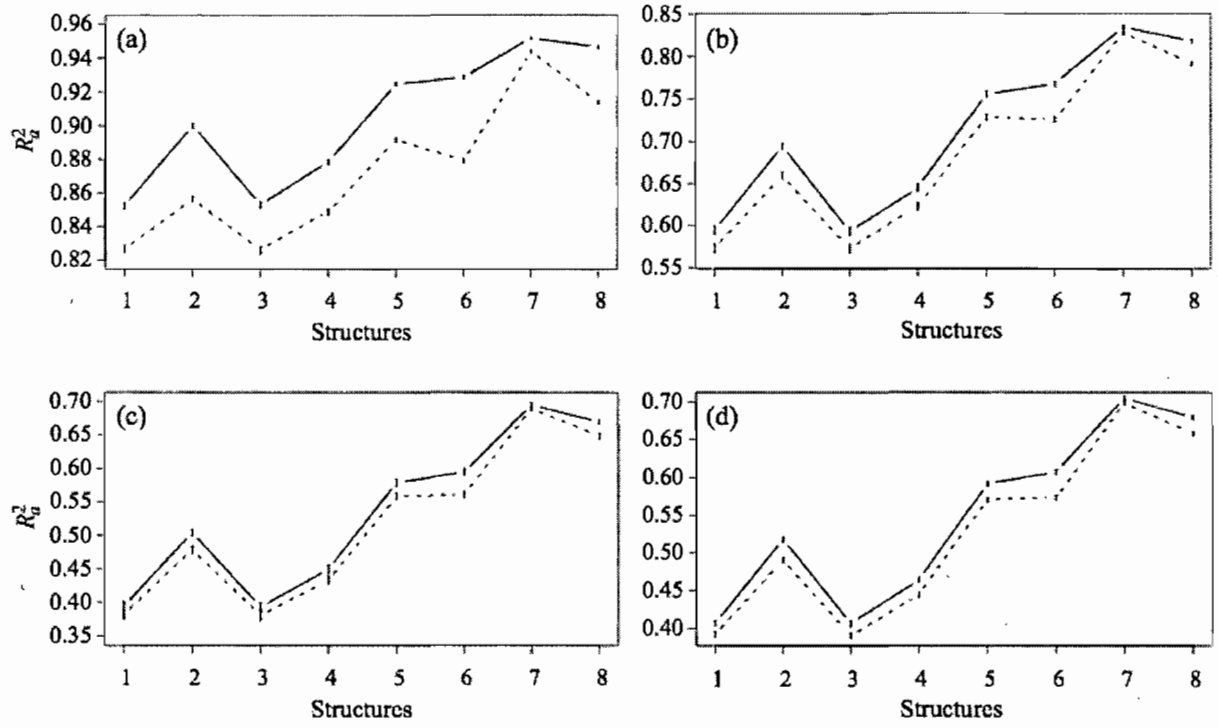


Fig. 4

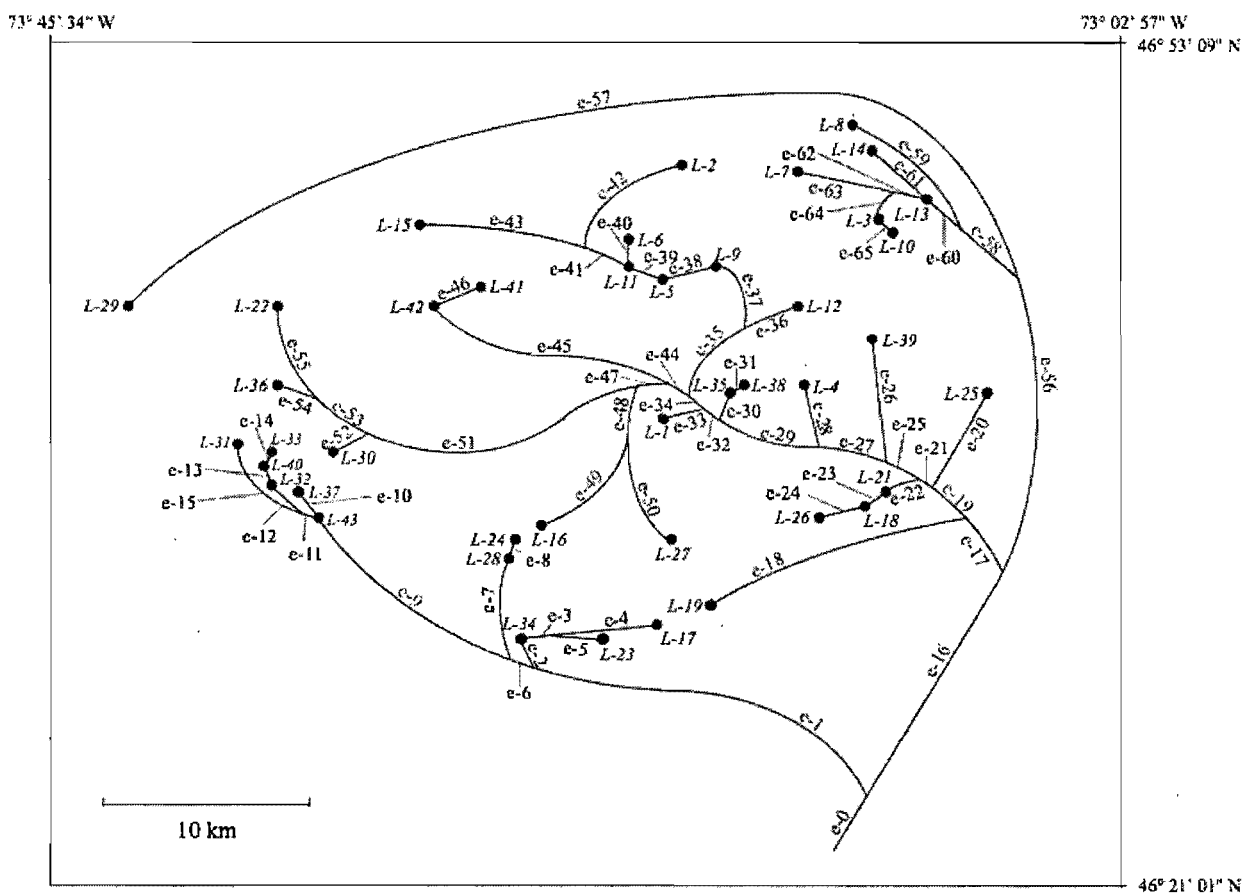


Fig. 5

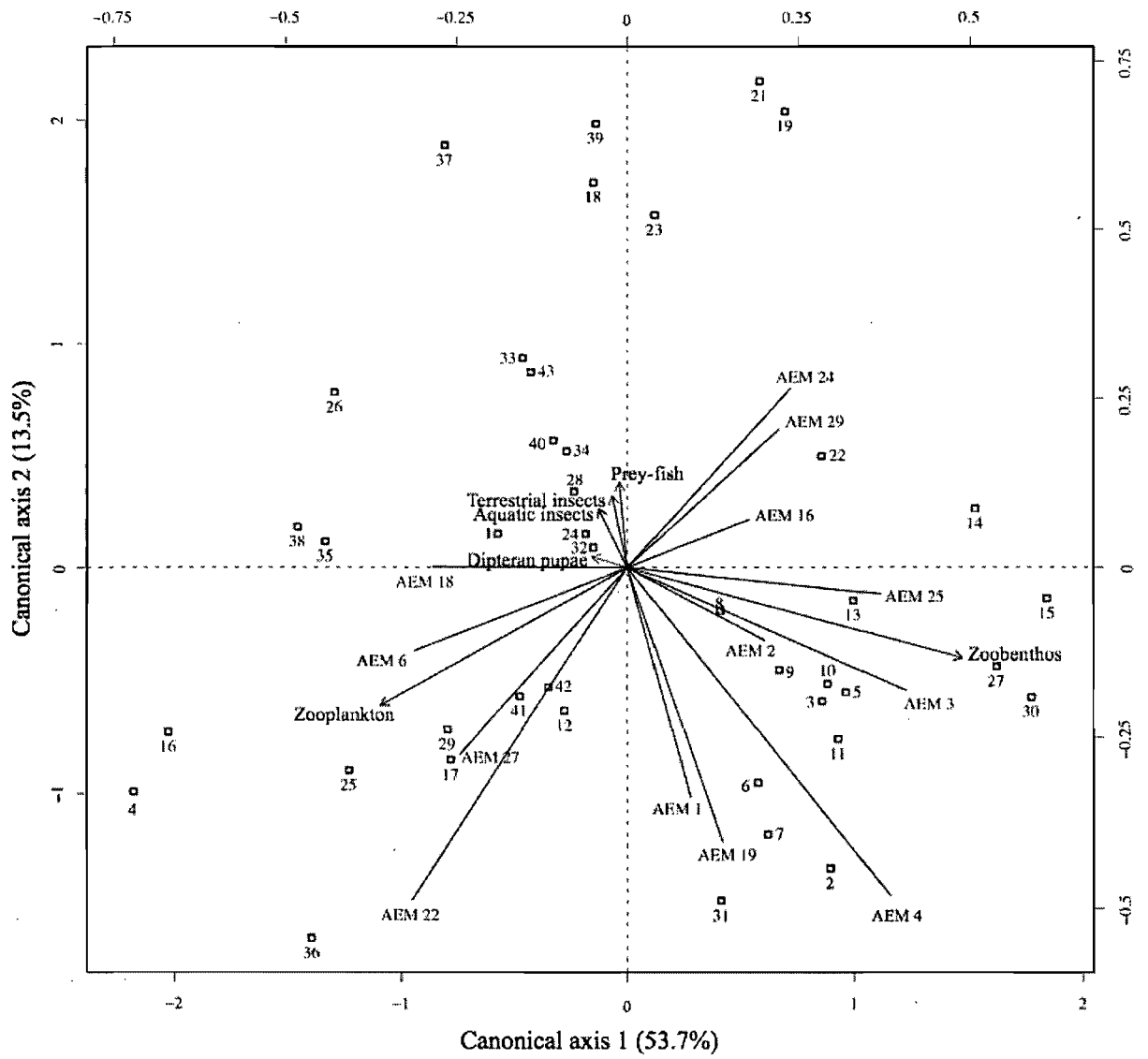


Fig. 6

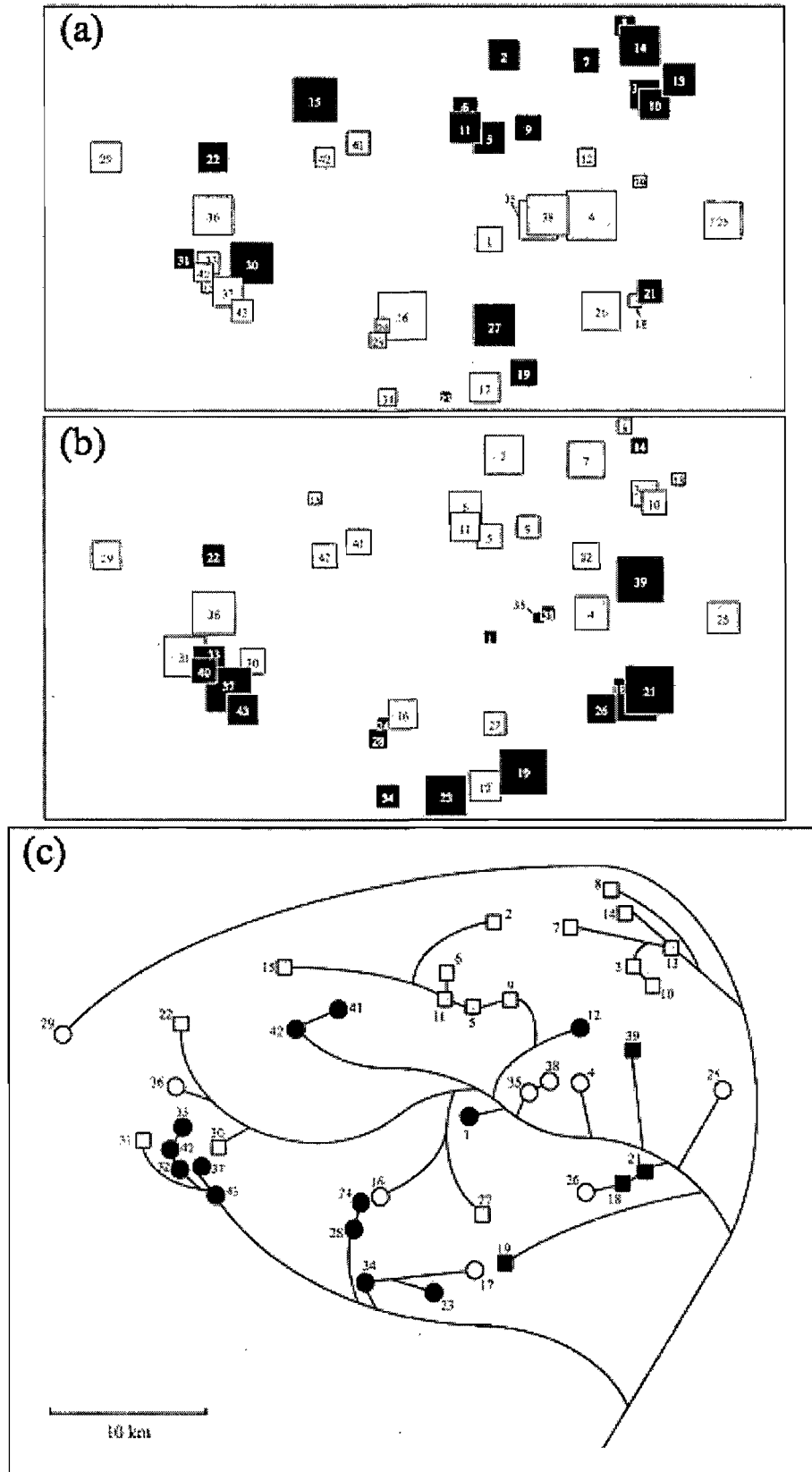


Fig. 7

CONCLUSION

Les deux chapitres de ce mémoire proposent des solutions à deux problèmes méthodologiques rencontrés en analyse spatiale des communautés d'espèces.

Le premier chapitre présente une solution élégante au problème de la sélection des variables spatiales orthogonales. Cette solution élimine les problèmes de la surexplication et de l'inflation importante de l'erreur de type I. De plus, cette méthode permet de conserver l'objectivité tant appréciée par les chercheurs qui utilisent les méthodes de sélection automatique classiques. Ce chapitre présente aussi une nouvelle approche permettant de tester des groupes de variables spatiales comportant $(n-1)$ variables, comme cela se produit fréquemment lorsque les variables spatiales servant à la modélisation sont créées dans le cadre des MEM. Il était jusqu'alors impossible de tester la signification statistique de tels jeux de variables parce que le nombre de variables était trop élevé. Ce développement permet aussi une meilleure interprétation des résultats obtenus lorsque la méthode de sélection progressive est utilisée avec des variables spatiales orthogonales puisque les résultats présentent maintenant une erreur de type I juste.

La nouvelle façon de créer des variables spatiales présentée dans le second chapitre de ce mémoire montre une tendance de plus en plus présente dans la littérature en écologie statistique, celle de créer des méthodes spécialisées pour un groupe de problèmes particuliers. Les AEM ont été développées pour des situations où la présence d'un processus spatial asymétrique est connue. Ce nouveau développement contribuera à une meilleure compréhension des processus régissant les communautés d'espèces simplement parce que la méthode statistique est plus adaptée aux données étudiées.

Ce mémoire présente des contributions à l'un des niveaux de la recherche scientifique, soit la méthodologie statistique. La nature est extrêmement complexe; chaque étape a son importance pour mieux la comprendre.

BIBLIOGRAPHIE (INTRODUCTION)

- Bergin T. M. 1992. Habitat selection by the western kingbird in western Nebraska: a hierarchical analysis. *The Condor*, 94:903-911.
- Borcard, D. and 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. and Tuosimoto, H., 2004. Dissecting the spatial structure of ecological data at multiple scales. *Ecology*, 85:1826-1832.
- Borcard, D., Legendre, P. and Drapeau, P., 1992. Partialling out the Spatial Component of Ecological Variation. *Ecology*, 73:1045-1055.
- Brind'Amour, A., Boisclair, D., Legendre, P. and Borcard, D., 2005. Multiscale spatial distribution of a littoral fish community in relation to environmental variables (vol 50, pg 465, 2005). *Limnology and Oceanography*, 50:465-479.
- Cooper S. D., Barmuta L., Sarnelle O., Kratz K., Diehl S. 1997. Quantifying spatial heterogeneity in streams. *Journal of the North American benthological society*, 16:174-188.
- Dray, S., Legendre, P. and Peres-Neto, P.R., 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecological Modelling*, 196:483-493.
- Duque, A.J., Duivenvoorden, J.F., Cavelier, J., Sanchez, M., Polania, C. and Leon, A., 2005. Ferns and Melastomataceae as indicators of vascular plant composition in rain forests of Colombian Amazonia. *Plant Ecology*, 178:1-13.
- Gauch, H.G., 1993. Prediction, Parsimony and Noise. *American Scientist*, 81:468-478.
- Gauch, H.G., 2003. *Scientific Method in Practice*. Cambridge University press, New York, 435 pages p.
- Griffith, D.A. and Peres-Neto, P.R., 2006. Spatial modeling in ecology: The flexibility of eigenfunction spatial analyses. *Ecology*, 87:2603-2613.

- Halpern, B.S. and Cottenie, K., 2007. Little evidence for climate effects on local-scale structure and dynamics of California kelp forest communities. *Global Change Biology*, 13:236-251.
- Huston, M.A., 1996. *Biological diversity: The coexistence of species on changing landscapes*. Cambridge University press, Cambridge, 681 p.
- Hutchinson, G.E., 1957. Concluding remarks. *Cold Spring Harbor Symposium on Quantitative Biology*, 22:415-427.
- Kolasa, J. and Rollo C. D. 1991. Introduction: The heterogeneity og heterogeneity - A glossary. In: J. Kolasa and S. T. A. Pickett (Editor), *Ecological heterogeneity*. Springer-Verlag, New York, pp. 1-23.
- Legendre, P., 1990. Quantitative methods and biogeographic analysis. In: D.J. Garbary and R.G. South (Editor), *Evolutionary biogeography of the marine algae of the North Atlantic*. Springer-Verlag, Berlin, pp. 9-34.
- Legendre, P. and Fortin, M.-J., 1989. Spatial pattern and ecological analysis. *Vegetatio*, 80:170-138.
- Planque B., Hays G. C., Ibanez F., Gamble J. C. 1997. Large scale spatial variations in the seasonal abundance of *Calanus finmarchicus*. *Deep-Sea Research I*, 44:315-326.
- Telford, R.J. and Birks, H.J.B., 2005. The secret assumption of transfer functions: problems with spatial autocorrelation in evaluating model performance. *Quaternary Science Reviews*, 24:2173-2179.