Global Ecology and Biogeography, (Global Ecol. Biogeogr.) (2010) **19**, 174–184

A Journal of Macroecology

Global Ecology and Biogeography

MACROECOLOGICAL METHODS

Estimating and controlling for spatial structure in the study of ecological communities

Pedro R. Peres-Neto[1]* and Pierre Legendre[2]

[1]*Département des Sciences Biologiques, Université du Québec à Montréal, C.P. 8888, Succursale Centre-ville, Montréal, Québec, Canada H3C 3P8,* [2]*Département de Sciences Biologiques, Université de Montréal, C.P. 6128, Succursale Centre-ville, Montréal, Québec, Canada H3C 3J7*

*Correspondence: Pedro R. Peres-Neto, Département des Sciences Biologiques, Université du Québec à Montréal, C.P. 8888, Succursale Centre-ville, Montréal, Québec, Canada H3C 3P8.
E-mail: peres-neto.pedro@uqam.ca

ABSTRACT

Aim Variation partitioning based on canonical analysis is the most commonly used analysis to investigate community patterns according to environmental and spatial predictors. Ecologists use this method in order to understand the pure contribution of the environment independent of space, and vice versa, as well as to control for inflated type I error in assessing the environmental component under spatial autocorrelation. Our goal is to use numerical simulations to compare how different spatial predictors and model selection procedures perform in assessing the importance of the spatial component and in controlling for type I error while testing environmental predictors.

Innovation We determine for the first time how the ability of commonly used (polynomial regressors) and novel methods based on eigenvector maps compare in the realm of spatial variation partitioning. We introduce a novel forward selection procedure to select spatial regressors for community analysis. Finally, we point out a number of issues that have not been previously considered about the joint explained variation between environment and space, which should be taken into account when reporting and testing the unique contributions of environment and space in patterning ecological communities.

Main conclusions In tests of species-environment relationships, spatial autocorrelation is known to inflate the level of type I error and make the tests of significance invalid. First, one must determine if the spatial component is significant using all spatial predictors (Moran's eigenvector maps). If it is, consider a model selection for the set of spatial predictors (an individual-species forward selection procedure is to be preferred) and use the environmental and selected spatial predictors in a partial regression or partial canonical analysis scheme. This is an effective way of controlling for type I error in such tests. Polynomial regressors do not provide tests with a correct level of type I error.

Keywords

Community analysis, eigenvector maps, model selection, spatial autocorrelation, spatial correlation, spatial predictors, variation partitioning.

INTRODUCTION

A major interest in ecology is to understand how factors such as local environment and landscape heterogeneity contribute to how regional pools of potential colonizer species are sorted into local communities across space. At least two challenges are encountered when studying community structure using distributional data. The first one is to understand how the species themselves are distributed in space. Due to spatially contagious processes, such as dispersal, differential mortality, social organization and species interactions, individual species tend to be spatially organized in space (Keitt *et al.*, 2002; Cottenie, 2005). Moreover, the environmental processes responsible for structuring communities may also be spatially organized, which in turn imposes a spatial structure, called *induced spatial dependence* (see Table 1 for definition), to communities. Here the challenge is to analyse the spatial component of communities in order to obtain clues regarding their origin and at which scales they act (e.g. Olden *et al.*, 2001; Borcard *et al.*, 2004; Diniz-Filho & Bini, 2005).

DOI: 10.1111/j.1466-8238.2009.00506.x

The second challenge is analytical in the sense that statistical tests may overestimate the importance of underlying ecological drivers, such as environmental factors (Legendre, 1993; Dale & Fortin, 2002), generating apparent species–environment concordance (Bivand, 1980; Legendre *et al.*, 2002; Dormann *et al.*, 2007). Here, the challenge is to filter out (control for) the variation due to spatial structures so that statistical tests may be applied properly under the assumption of independence (Griffith & Peres-Neto, 2006; Dormann *et al.*, 2007; Hawkins *et al.*, 2007). Taken together, mathematical representations of spatial relationships among the study sites may be seen as predictors when the goal is to understand how communities are organized in space ('spatial legacy'), or as covariables when the goal is to filter out spatial variation ('spatial nuisance'; see Table 1 for definitions). Note that throughout this paper we refer to spatial autocorrelation only to the spatial structure in species distributions that are the result of intrinsic population or community dynamics (following Fortin & Dale, 2005; see Table 1 for definition) and not by induced spatial dependence. The terms spatial structure, spatial dependence and spatial correlation, however, are used interchangeably (see Table 1).

Canonical analyses such as redundancy analysis (RDA; Rao, 1964) and canonical correspondence analysis (CCA; ter Braak, 1986) provide the means of conducting direct explanatory analyses in which ecological communities can be studied with respect to their relationships with ecological drivers. In the realm of canonical analysis, variation partitioning (Borcard *et al.*, 1992; Diniz-Filho & Bini, 2005; Legendre *et al.*, 2005; Peres-Neto *et al.*, 2006) is routinely used in ecological analysis to: (1) estimate the contribution of spatial variation in structuring communities (spatial legacy), and (2) filter out the effects of spatial correlation when testing for the importance of ecological factors such as environmental predictors (spatial nuisance). There are different ways of controlling for spatial correlation in single-species models (see Dormann *et al.*, 2007, for a review). However, for the case of multi-species data, variation partitioning is the most commonly used technique in community analyses. Despite its broad use, to date there has been no attempt to assess whether the variation partitioning technique does offer a solution to the problem of spatial correlation, or how the different types of spatial predictors that can be used in the technique compare. With this shortcoming in mind, the goal of this study is twofold: (1) compare the success of different spatial techniques in estimating the spatial component of ecological communities, and (2) assess whether these techniques are successful in removing the statistical bias due to spatial correlation when testing for the importance of environmental drivers under variation partitioning. We also discuss several issues revolving around spatial correlation and community analysis, offering a viewpoint that should be useful when exploring and interpreting patterns in ecological communities produced by environmental and spatial variation.

INNOVATION

Methods

Variation partitioning and spatial predictors

Canonical analyses such as RDA and CCA are methods that extend multiple linear regression, which has a single response y and multiple predictors x (e.g. several environmental predictors), to multivariate linear regression involving multiple response variables $\boldsymbol{Y}$ (e.g. several species) and a common matrix of predictors $\boldsymbol{X}$. Variation partitioning for RDA or CCA using one set of response variables ($\boldsymbol{Y}$, species) and two sets of predictors ($\boldsymbol{X}$, environment and $\boldsymbol{W}$, space) is straightforward (Fig. 1; see Anderson & Gribble, 1998 for an extension to multiple matrices).

When both species and environment are spatially structured but independently distributed (i.e. uncorrelated), tests solely based on the environmental component (i.e. component [*ab*] in Fig. 1) are biased (they have type I error rates that exceed the significance level: Dutilleul, 1993; Legendre *et al.*, 2002). If only the species data are spatially structured and the environment not, or vice versa, then tests are not biased (Dutilleul, 1993). In

Table 1 Definitions regarding the spatial terminology used in the paper.

Terminology	Definition
Spatial structure, spatial dependence, spatial correlation or spatial signature (used interchangeably throughout the text)	Any non-random organization across space in either species distributions (or their communities) or environmental processes
Spatial autocorrelation (or non-induced spatial dependence)	Spatial structure due to the dynamics of the species (or their communities) themselves (e.g. via dispersal)
Induced spatial dependence	Spatial structure in species distributions (or their communities) indirectly induced by the environment and not by autocorrelation
Spatial legacy	Spatial structure in species distributions generated by spatial autocorrelation or induced by exogenous processes (e.g. environmental factors). This structure pertains to the variation found in fraction [*c*] (see Fig. 1)
Spatial nuisance	The common spatial structure in species distributions and environmental processes that increases type I errors and can potentially affect model estimation. This structure pertains to the variation found in fraction [*b*] (see Fig. 1 and the Discussion section for more details).

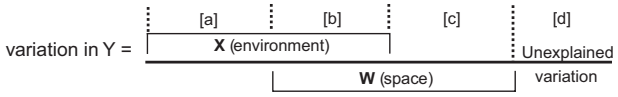

Figure 1 Variation partitioning scheme of a response variable $\boldsymbol{Y}$ between two sets of predictors, $\boldsymbol{X}$ (e.g. environmental factors) and $\boldsymbol{W}$ (e.g. spatial predictors). The total variation in $\boldsymbol{Y}$ is partitioned into seven components as follows: (1) calculate fraction $[a + b + c]$ based on both sets of predictor matrices $[\boldsymbol{X},\boldsymbol{W}]$ ($[a + b + c] = R^2_{Y|[X,W]}$); (2) calculate fraction $[a + b]$ based on matrix $\boldsymbol{X}$ ($[a + b] = R^2_{Y|X}$); (3) calculate fraction $[b + c]$ based on matrix $\boldsymbol{W}$ ($[b + c] = R^2_{Y|W}$); (4) the unique fraction of variation explained by $\boldsymbol{X}$: $[a] = [a + b + c] - [b + c]$; (5) the unique fraction of variation explained by $\boldsymbol{W}$: $[c] = [a + b + c] - [a + b]$; (6) the common fraction of variation shared by $\boldsymbol{X}$ and $\boldsymbol{W}$: $[b] = [a + b + c] - [a] - [c]$; and (7) the residual fraction of variation not explained by $\boldsymbol{X}$ and $\boldsymbol{W}$: $[d] = 1 - [a + b + c]$. The fractions are actually estimated by adjusted R^2 statistics (Peres-Neto *et al.*, 2006).

the presence of spatial correlation, observations are not independent and therefore the effective number of degrees of freedom in the sample is smaller than the one expected based on the number of observations used in the analysis, thus generating confidence limits that are spuriously narrower. This makes tests less conservative by increasing type I errors and may lead to incorrect conclusions about the effect of the environment on species. In order words, the effective type I error is greater than the significance level (or alpha-level) of the test, hence generating 'apparent species–environment concordance'. Variation partitioning offers a natural way of not only dealing with this problem but also assessing the importance of spatial correlation in community structure. By considering the spatial and environmental predictors together, one aims at discounting the spatial variation in the environmental data so that the test based on the corrected fraction (i.e. test of fraction $[a]$) has a type I error equal to the pre-established significance level alpha. Moreover, one can assess the amount of variation that is due to space alone independently of the environmental factors considered (but see discussion section) by testing fraction $[c]$ (i.e. spatial variation independent of space; see Fig. 1).

Two methods for generating spatial predictors to compose matrix $\boldsymbol{W}$ were considered here due to their extensive use in variation partitioning applied to ecological communities. The most common approach is to generate trend surface polynomials based on the geographic coordinates of the study sites (e.g. Gittins, 1985; Wartenberg, 1985; Borcard *et al.*, 1992; Legendre & Legendre, 1998; Lichstein *et al.*, 2002; and references therein). This method has been criticized because the spatial structures generated are rather simple and global, such as a gradient, a single wave or a saddle (Legendre & Legendre, 1998). We generated spatial predictors based on a third-degree polynomial that were orthogonalized using a principal component decomposition. Note that the orthogonalization does not affect polynomial performance but increases the computational speed of model selection procedures (see Model Selection section). The other approach was based on an eigenvector decomposition of connectivity matrices, namely Moran's eigenvector maps (MEM) by Dray *et al.* (2006) (see Griffith & Peres-Neto, 2006, for a review and description of other related methods). MEMs are relatively new but their use is becoming quite common, even in variation partitioning (e.g. Diniz-Filho & Bini, 2005; Legendre *et al.*, 2009; Yamanaka *et al.*, 2009). Spatial eigenvector mapping is based on the idea that the spatial relationships among data points can be translated into explanatory variables, which capture spatial effects at different spatial scales. Eigenvectors from these connectivity matrices represent the decompositions of Moran's I statistic into all mutually orthogonal maps that can be generated from a given connectivity matrix (Griffith & Peres-Neto, 2006). There are several possibilities for building connectivity matrices and we followed the implementation suggested in Borcard & Legendre (2002) based on the maximum distance that keeps all sites linked, which is calculated on the basis of a minimum spanning tree. Eigenvectors having associated eigenvalues that are positive represent positive spatial correlation, whereas eigenvectors having negative eigenvalues represent negative spatial correlation. MEMs are generated as linearly independent by default and hence no further orthogonalization is required, as in the case of polynomial regressors. A more complete description of MEMs and other related methods (e.g. PCNM; Borcard & Legendre, 2002) is offered in the additional Supporting Information (Appendix S1).

Data simulation and comparisons

The model for simulating species distributions based on spatial and environmental predictors followed Legendre *et al.* (2002, 2005):

$$\boldsymbol{S}_j = \boldsymbol{SA}_j + \beta\boldsymbol{E}_j + \boldsymbol{\varepsilon}_j \quad (1)$$

where $\boldsymbol{S}_j$ is a (10,000-site × 1) vector representing the distribution of the jth species, $\boldsymbol{SA}_j$ is a (10,000-site × 1) vector containing $N(0,1)$ deviates that are conditioned to be spatially autocorrelated, $\boldsymbol{E}_j$ is a (10,000-site × 1 environmental variable) matrix containing $N(0,1)$ deviates that in certain simulation scenarios were also conditioned to be spatially autocorrelated, β is the strength of the environmental component, and $\boldsymbol{\varepsilon}_j$ is a (10,000-site × 1) vector containing $N(0,1)$ deviates for species j. For instance, when both the species and environmental data were generated independently ($\beta = 0$) and both were spatially conditioned we used these data to estimate how well the different spatial predictors performed in controlling inflated type I error rates in tests of the importance of the environmental drivers. Spatial structure in $\boldsymbol{E}_j$ and $\boldsymbol{SA}_j$ was induced by a conditional sequential Gaussian algorithm (implemented as in Deutsch & Journel, 1992) to generate spatial realizations according to a spherical variogram model with a specified range (arbitrary grid units) on a 100 × 100 lattice, hence 10,000 sites. In all simulations, $\boldsymbol{S}$ contained 10 species (i.e. $\boldsymbol{S}$ is a 10,000 × 10 matrix). The first five species (columns in $\boldsymbol{S}$) were simulated always without spatial correlation and environmental control ($\beta = 0$; species were just normally distributed $N(0,1)$ variables), whereas the remaining species were generated with or without an environ-

Table 2 Simulation scenarios used in this study.

Scenarios	β	Variogram range		Type of assessment					Result
		Environment	Species	[*abc*]	[*ab*]	[*bc*]	[*a*]	[*c*]	
1	0	*N*(0,1)	–	Type I error	Type I error	Type I error	Type I error	Type I error	Fig. 2
2	0	*N*(0,1)	30	Power	Type I error	Power	Type I error	Power	Fig. 3(a)
3	0	30	30	Power	Type I error	Power	Type I error	Power	Fig. 3(b)
4	0	50	50	Power	Type I error	Power	Type I error	Power	Fig. 3(c)
5	0	80	80	Power	Type I error	Power	Type I error	Power	Fig. 3(d)
6	0.25	15	15	Power	Power	Power	Power	Power	Fig. 4(a)
7	0.50	15	15	Power	Power	Power	Power	Power	Fig. 4(b)
8	0.25	30	30	Power	Power	Power	Power	Power	Fig. 4(c)
9	0.50	30	30	Power	Power	Power	Power	Power	Fig. 4(d)

Each scenario represents a combination of the strength (slope) of the environmental component β (equation 1), the variogram range parameter values used to generate spatial structures in the environment (***E*** in equation 1) and the species (***SA*** in equation 1). Non-spatialized environment was generated using *N*(0,1) as ***E*** in equation 1 and non-autocorrelated species were produced by eliminating ***SA*** in equation 1. The type of assessment (power versus type I error) depends on the component ([*abc*] = environment + space, [*ab*] = environment confounded with space or [*bc*] = space confounded with environment) and fraction ([*a*] = pure environment or [*c*] = pure space). See Fig. 1 for a definition and calculation of components and fractions. The 'Result' column indicates the figure in which the rejection rates for each scenario and respective components and fractions are reported.

mental control or spatial correlation depending on the simulation scenario (see Table 2). For a given variogram range, all species were generated with the same range; however, each ***SA*** vector contained an independent realization based on separate runs of the conditional sequential Gaussian algorithm. Each species generated with an environmental control was related to a single environmental variable (i.e. five environmentally controlled species in ***S*** and five environmental variables in ***E***). Scenarios were established by manipulating β (0, 0.25 or 0.50) and the strength of the spatial correlation component in the species' ***SA*** vector and environmental data (i.e. variogram range as 15, 30, 50 or 80). Table 2 presents the different scenarios and what they represent in terms of assessment (i.e. type I error versus statistical power of the test). For each of these scenarios, we produced 1000 sample data sets and comparisons between methods were performed on the basis of their type I error and power rates. Type I error rates were estimated as the fraction of tests (over 1000 sample data sets) that erroneously rejected the null hypothesis when was set as true (i.e. when β or ***SA*** was set to zero; see Table 2); whereas power rates were estimated as the fraction of tests (over 1000 sample data sets) that correctly rejected the null hypothesis when was set as false (i.e. when β or ***SA*** was set as different from zero; see Table 2). In all cases, an alpha of 0.05 was applied. Also, we only present results for the cases where both species and environment present similar ranges. Results based on different ranges in species and environment provided similar results in terms of contrast of the two spatial methods and will be not presented for brevity.

As generated above, ***S*** contains normally distributed data. However, ecologists are most interested in the case where response variables are counts of species abundances or presence/absence data, which are generally overdispersed and zero-inflated (Martin *et al.*, 2005). In order to generate discrete and zero-inflated data, we transformed a simulated population matrix ***S*** as follows: $\boldsymbol{S}' = [s'_{ij}] = \exp(1.5\boldsymbol{S}_{\text{std}})$. Once generated, matrix ***S*** was first standardized (zero mean and unit variance) into $\boldsymbol{S}_{\text{std}}$ so that a standard deviation of 1.5 could be then forced. The values s'_{ij} were then rounded to the lower integer to generate ***S′***, which contained roughly 50% zeros; standard deviations different from 1.5 would generate different numbers of zeros in ***S′***. In addition, we also considered the case of presence/absence species data matrices where abundance values larger than zero in ***S′*** were transformed to 1s.

From each species matrix ***S′*** (10,000 × 10) and environmental matrix ***E*** (10,000 × 5), we sampled 100 observations using a square grid design with 10 × 10 sampling points, with a spacing of 10 units between adjacent neighbour sites to compose the sample species matrix ***Y*** and environmental matrix ***X***. ***Y*** was Hellinger-transformed prior to analysis (see Peres-Neto *et al.*, 2006, for the properties of this distance regarding variation partitioning). This transformation involves a square-root transformation and can normalize Poisson or Poisson-like data (e.g. abundance data). Spatial predictors (i.e. polynomials and MEMs) were constructed according to the spatial coordinates of the 10 × 10 sampling grid. For each scenario, based on combinations of the strength of the environmental gradient (β) and strength of the spatial correlation (variogram range) in the species and environmental data, 1000 data sets (100 × 100 map and corresponding 10 × 10 sample of points) were generated.

Model selection for spatial predictors

Statistical procedures for testing overall and unique contributions of environment and spatial predictors in variation partitioning (Fig. 1) are commonly used and well established in the literature (see Legendre & Legendre, 1998, pp. 608–612) and can be performed by software such as the Canoco 4.5 program (ter Braak & Šmilauer, 2002) and the R language (the 'varpart'

function of the 'vegan' library by Oksanen *et al.*, 2008). Spatial predictors were generated and selection procedures were performed in MATLAB (release 2007a, The MathWorks); functions are available from the senior author.

Based on our previous experiences with variation partitioning (Legendre *et al.*, 2005; Peres-Neto *et al.*, 2006), we anticipated that the number of spatial predictors in the model would influence the power of the test to detect unique and significant environmental contributions (i.e. fraction [*a*], Fig. 1) due to a decrease in degrees of freedom in the overall model. Since the number of spatial predictors can be quite large (for instance, in our 10 × 10 lattice containing 100 points, 32 MEMs and 9 polynomials were considered as spatial predictors), it is essential to apply a model selection procedure in order to reduce the number of spatial predictors so that significant unique environmental fractions have a greater chance of being identified. However, given that model selection procedures may inflate type I error rates (Wilkinson & Dallal, 1981; Blanchet *et al.*, 2008; Mundry & Nunn, 2009; that fact was also observed in early versions of our simulations), we first tested the significance of the spatial model (i.e. component [*bc*] in Fig. 1) with all spatial predictors for a given type of spatial regressor (i.e. polynomial or MEMs). If significant, we then proceeded with forward selection to determine an adequate number of predictors describing spatial variation. That subset was then used in all tests for a particular set of simulated data.

In this study, only the positively autocorrelated spatial regressors were considered as spatial candidates since the simulated spatial patterns were all positively autocorrelated. The most common model selection procedure used in canonical analyses is forward selection. Note that since the spatial predictors used here are orthogonal to one another, no difference is to be expected between forward selection and a number of other procedures (e.g. backward selection and stepwise procedures; McQuarrie & Tsai, 1999). Due to their orthogonality, the forward selection procedure is much faster as the order of entrance in the model can be determined a priori by the amount of explained variation of a given spatial predictor. The complete description of the procedure is presented as Supporting Information (Appendix S2). Note, however, that in the case of polynomials we also performed all the simulations using the original polynomials (i.e. not their principal components) but they were similar in performance, and hence only results based on the principal components are presented.

In initial simulations, we noticed that the standard forward selection procedure did not correctly control type I error rates due to spatial correlation (see Results section). This result is due to the fact that the forward selection for multiple species is based on the principle of parsimony where only spatial predictors that contribute to at least a few species are likely to be detected. This problem is somewhat akin to analysis of variance where a large number of sample means are considered; if all sample means are equal, except one (or a small number of samples), the power of the test to detect this difference is low when compared with a smaller number of sample means. Therefore, we decided to apply the forward selection procedure to each species separately. The spatial matrix $\boldsymbol{W}$ was the union of the predictors resulting from each separate model. For instance, consider two species A and B. Assume that for species A, predictors [4,5] were retained whereas for species B predictors [5,9] were retained; then matrix $\boldsymbol{W}$ would be composed of predictors [4,5,9].

RESULTS

Results are presented in the form of rejection rates over 1000 sample tests. Table 2 presents the simulation scenarios, which were based on a combination of the strength of the environmental (β) and spatial gradients (variogram range) on the species. For instance, when β was set to zero but the environmental and species spatial gradients were generated with a range of 15 (or greater), we could then assess: (1) how spatial dependence induced a bias in the test of the environmental component [*ab*], and (2) the performance of the different spatial predictors in controlling this bias while testing fraction [*a*]. Table 2 also indicates the figures in which the results for each scenario are presented. Note, again, that we first tested the overall significance of the spatial predictors (i.e. component [*bc*]) with all spatial predictors for a given type of spatial regressor; if significant, we then conducted a forward selection to retain the spatial predictors to be used in testing the components and fractions involving the environmental and spatial predictors (i.e. [*abc*], [*a*] and [*c*]). For each type of predictor we also present the results for analyses based on all spatial predictors (i.e. all 32 MEMs and 9 polynomials).

The first scenario was based purely on random data, where species, environment and the species spatial component were randomly $N(0,1)$ generated. In this case, testing for components and fractions presented correct type I error in all cases (Fig. 2), showing that the two-step procedure, in which we first test all predictors and then run the model selection, is an effective way of controlling the inflated type I errors in model selection as previously shown in the literature (see Methods).

Scenario 2 (Table 2, Fig. 3a) serves to show that when spatial dependence is present only in the species [i.e. $\boldsymbol{E} = N(0,1)$ and range of $\boldsymbol{SA} = 15$ for the species], tests for the environmental component [*ab*] are not biased. Results for other $\boldsymbol{SA}$ ranges [i.e. $\boldsymbol{E} = N(0,1)$ and range of $\boldsymbol{SA} = 30$ and 50] provided comparable results (i.e. correct levels of type I error) but are not reported here for brevity. However, when both the species and environment are spatially structured but without any contribution of the environment to the species variation (scenarios 3–5, i.e. Fig. 3b–d), the type I error of the environmental component [*ab*] steadily increases with the strength of the spatial dependence (i.e. range). Regarding the control over elevated type I error rates (i.e. tests of fraction [*a*]), polynomial regressors, which are the most commonly used in variation partitioning and other techniques, fail to provide a valid test (i.e. rejection rates for fraction [*a*] are greater than the pre-established alpha of 0.05). Note, however, that their effectiveness (reduced type I error rates) increases with the range of the variograms controlling the spatial structure (Fig. 3b–d). The standard forward selection for both MEM regressors, although more robust, also

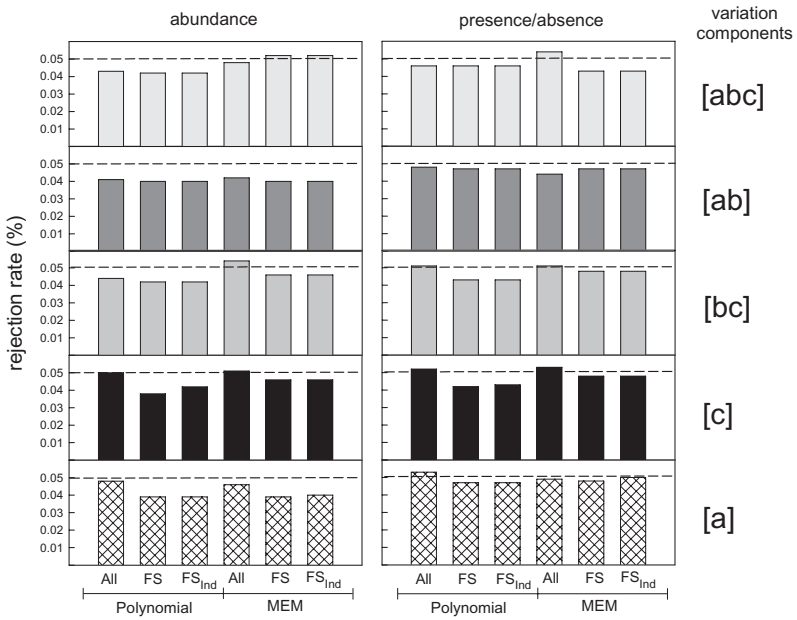

Figure 2 Type I error rates of the tests when no spatial component (i.e. ***SA*** in equation 1) and no effect of the environment on the species data were present. Rejection rates measured as the proportion of rejections ($\alpha = 0.05$) per 1000 sample tests for scenario 1 (see Table 2) for the polynomial and Moran's eigenvector maps (MEM) spatial predictors according to model selection procedures and fractions and components of variation. Expected nominal type I error rates (0.05) are represented by a dashed line. Here all rejection rates represent type I error rates as species, environment and the species spatial component were randomly [$N(0,1)$] generated. 'All' indicates that all spatial variables generated for each method were used, FS denotes forward selection. and FS_{Ind} represents individual species forward selection. Note that all models and selection procedures have correct nominal type I error rates when all component of variation are random. Variation components and fractions (i.e. [*abc*], [*ab*], [*bc*], [*a*] and [*c*]) are defined in Fig. 1.

fails in controlling high type I error rates in the tests of fraction [*a*]. Nevertheless, correct control over correlation (i.e. inflated type I errors) is effective either using all MEM regressors or the individual species selection model.

The results from scenario 2 (Fig. 3a) also indicates that MEM presents the greatest power to detect spatial structure in species distributions (fraction [*c*] and components [*abc*] and [*bc*]) at small spatial scales (range = 15) whereas all methods presented similar power for medium to larger spatial structures (ranges 30, 50 and 80; Fig. 3b–d). Given that all methods have similar power to detect large spatial structures, we further compared their power to detect unique environmental (i.e. fraction [*a*]) and spatial (i.e. fraction [*c*]) contributions only based on small-scale and medium spatial structures (i.e. range = 15 and 30; Fig. 4a–d). The results clearly indicate that MEM presents the greatest power to detect spatial structures ([*c*], [*bc*] and [*abc*]), especially based on the individual species selection model. Power for testing fraction [*a*] was slightly greater for the polynomial regressors for the smaller scale (range = 15) but much greater for medium scales (range = 30), which is of course offset by the fact that this method presents increased type I error rates (Fig. 3). Using all predictors severely decreases the power of MEM.

DISCUSSION

Our goal was twofold. Firstly, to test whether variation partitioning applied to canonical analysis is an appropriate method for detecting spatial structures in ecological data (presence–absence and abundance) and controlling for inflated type I errors when testing for the importance of the environmental component in driving species distribution. Secondly, we wanted to provide guidance regarding the type of spatial predictor to use in variation partitioning. Since variation partitioning is the most widely used technique in determining the importance of environmental and spatial variation in structuring ecological communities, this assessment is timely.

An important consideration is whether a variation partitioning scheme is necessary for the data at hand in the first place. If the species data only are autocorrelated while the environmental data are spatially independent, or vice versa, a variation partitioning scheme per se perhaps should not be considered. In that case, separate models relating community to environment (component [*ab*]) and community to space (component [*bc*]) should be used; the reason is that the tests of the unique fractions ([*a*] and [*c*]) is penalized by the loss of unnecessary degrees of freedom by considering space and environment jointly.

Our results clearly indicate that among the methods considered here, polynomial spatial regression, which is the most widely used method for incorporating spatial variation in variation partitioning, does not control for the bias and presents the highest type I error rates when testing the environmental contribution. Because of its common use, we considered a third-degree polynomial, though other orders could have been used (e.g. Rowe & Lidgard, 2009). Note that in practice, polynomials

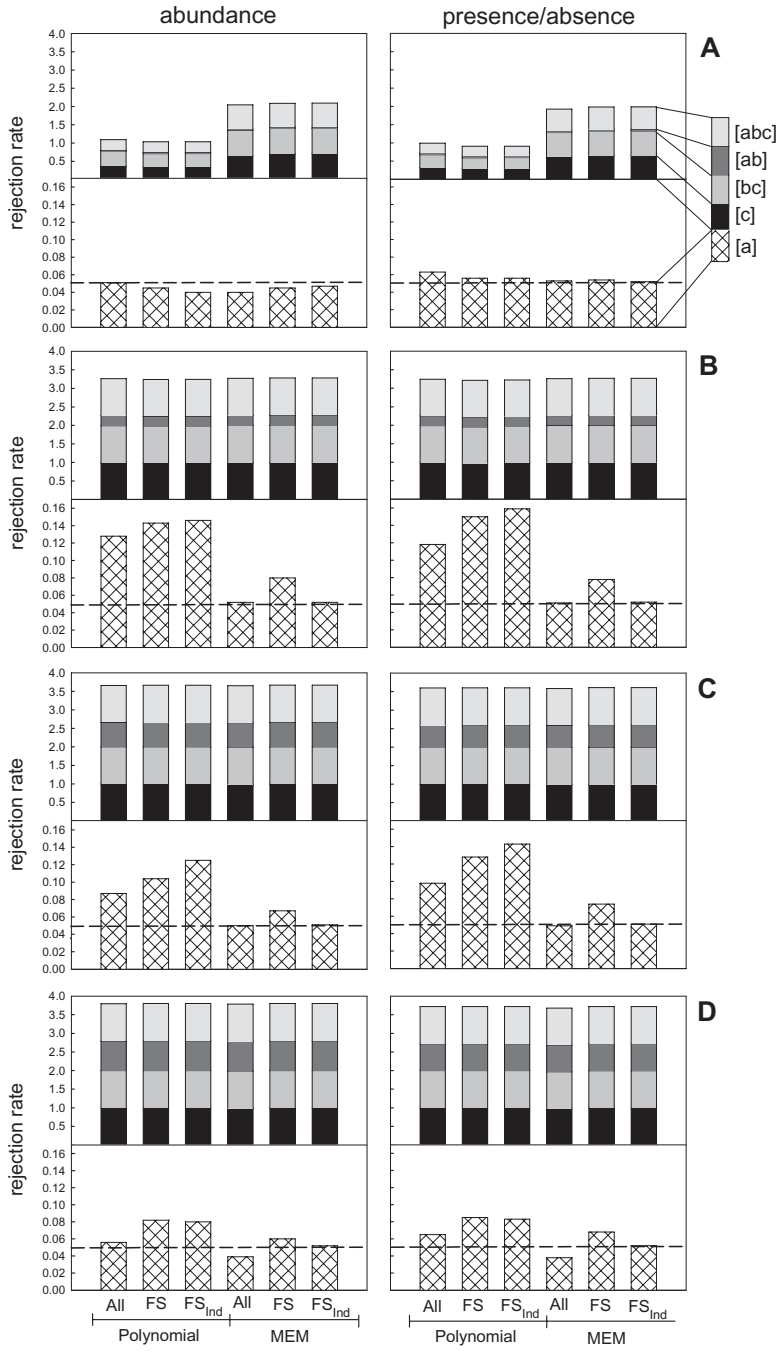

Figure 3 Results of the simulations in the presence of a spatial component in the species data (i.e. ***SA*** in equation 1), without any environmental effect (i.e. ***E*** in equation 1) on the species. Rejection rates measured as the proportion of rejections ($\alpha = 0.05$) per 1000 sample tests for the polynomial and Moran's eigenvector maps (MEM) spatial predictors according to model selection procedures and fractions and components of variation. Panels (a) to (d) represent scenarios 2 to 5, respectively (see Table 2). Spatial dependence (variogram range) increases from panel (a) to (d). The lower graphs within each panel represent type I error rates for fraction [*a*] (i.e. unique fraction due to environment, independent of space). Expected nominal type I error rates (0.05) for fraction [*a*] are represented by a dashed line. The rejection rates for the other components and fractions (i.e. [*abc*], [*ab*], [*bc*] and [*c*]) are shown as stacked bars with each height representing the rejection rate of a component or fraction (respective bars are shown in panel (a), right side). Variation components and fractions are defined in Fig. 1. Here, all rejection rates for component [*ab*] and fraction [*a*] represent type I error rates (i.e. environmental influence is set to zero but is spatially structured) whereas for components [*abc*], [*bc*] and fraction [*c*], they represent power estimates (i.e. species are spatially autocorrelated). 'All' indicates that all spatial variables generated for each method were used, FS denotes forward selection and FS_{Ind} represents the individual species forward selection. Left: species abundance data; right: species presence–absence data. Polynomial regressors have low performance in controlling type I error rates (i.e. tests of fraction [*a*]), especially when the variogram range is small. The standard forward selection fails in controlling high type I error rates in the tests of fraction [*a*].

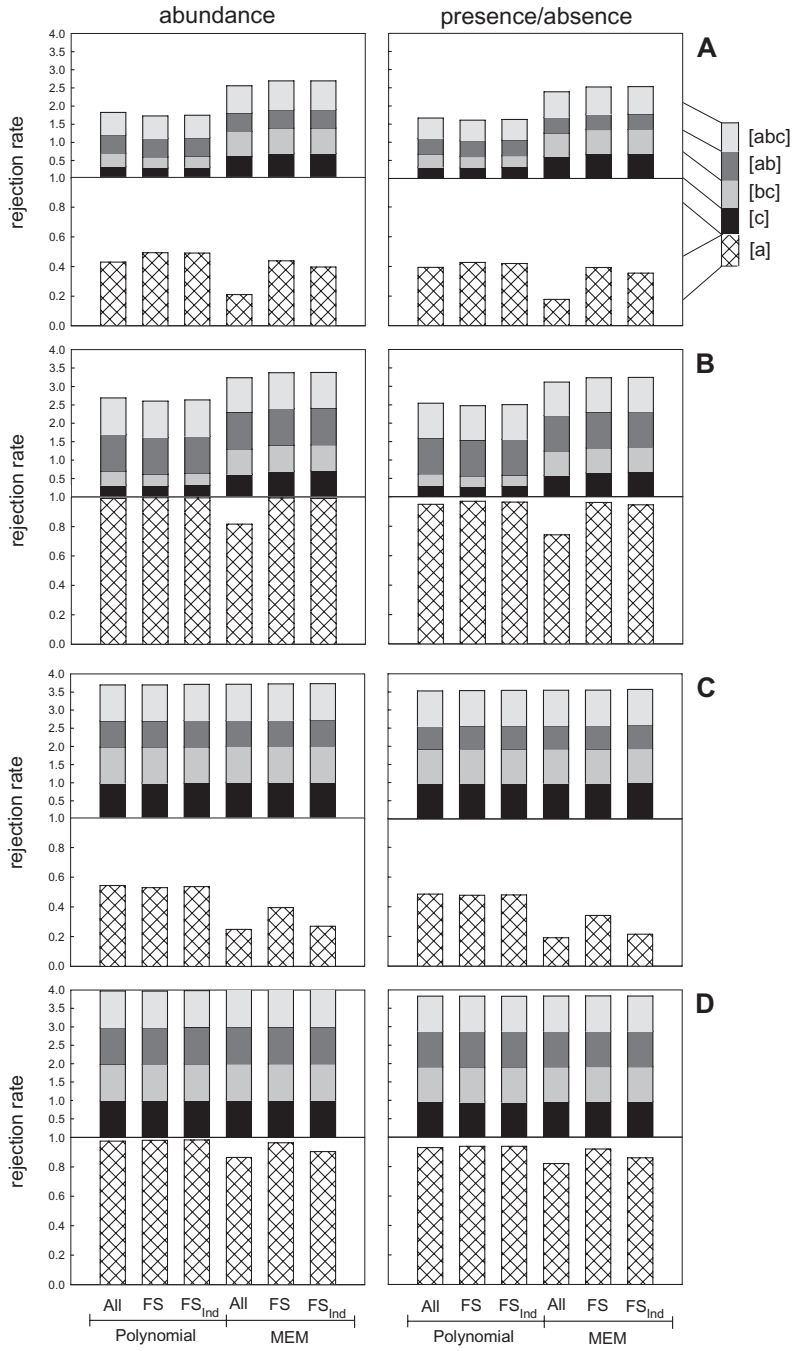

Figure 4 Results of the simulations in the presence of a spatial component in the species (i.e. ***SA*** in equation 1) and environmental data (i.e. ***E*** in equation 1), with influence of the environment on the species. Rejection rates measured as the proportion of rejections ($\alpha = 0.05$) per 1000 sample tests for the polynomial and Moran's eigenvector maps (MEM) spatial predictors according to model selection procedures and fractions and components of variation. Panels (a) to (d) represent scenarios 6 to 9 respectively (see Table 2). Spatial dependence (variogram range) is fixed at 15 (panels a and b) and 30 (panels c and d). The lower graphs within each panel represent type I error rates for fraction [*a*] (i.e. the unique fraction due to the environment, independent of space). The rejection rates for the other components and fractions (i.e. [*abc*], [*ab*], [*bc*] and [*c*]) are shown as stacked bars with each height representing the rejection rate of a component or fraction (respective bars are shown in panel a, right side). Variation components and fractions are defined in Fig. 1. Here, all rejection rates represent power estimates for all components and fractions (i.e. species and environment are spatially autocorrelated and the environmental contribution is greater than zero; for panels a and c $\beta = 0.25$ and for panels b and d $\beta = 0.50$). 'All' indicates that all spatial variables generated for each method were used, FS denotes forward selection and FS_{Ind} represents the individual species forward selection. MEM presents the greatest power in detecting spatial structure in species distributions (fraction [*c*] and components [*abc*] and [*bc*]) at small spatial scales (range = 15) whereas all methods presented similar power for medium to larger spatial structures (ranges 30, 50 and 80; Fig. 3b–d).

with greater degrees require very regular spacing among sampling units (Unwin, 1978). Their performance was not surprising, since we know that polynomials are more effective in modelling large-scale spatial structures. The greatest problem is that polynomials are not capable of characterizing regular cyclic changes in direction. The strength of MEM is that it generates these cyclic changes at multiple scales that when combined and weighted by regression coefficients are capable of representing very complex spatial patterns (see fig. 2 in Griffith & Peres-Neto, 2006).

We also demonstrated the importance of model selection in the case of spatial predictors. If all spatial predictors are used, the power to detect the environmental contribution (fraction [*a*]) is much reduced due to the reduced number of degrees of freedom in the denominator of the *F*-statistic. Our results also indicate that the standard forward selection model implemented in statistical packages that conduct canonical analyses does not necessarily provide an appropriate procedure for controlling inflated type I error rates. The individual species selection model provides a better solution for the problem and also opens up the possibility of using other techniques to control for spatial correlation only available for individual species, such as autoregressive models. Note that each species was generated independently with regard to their environmental control and intrinsic spatial dependence; although these are extreme cases, hence our decision to generate data under such a structure, species in nature will have various types of similarities regarding their environmental and spatial associations. The independent species spatial model selection introduced here, however, would also apply in these cases, and as the spatial communality increases among species, the individual selection approach would approach similar results to those of the overall standard forward selection approach.

Most ecologists are familiar with the idea that spatial correlation is a source of 'nuisance' in determining the importance of ecological drivers such as environment (Legendre, 1993). The spatial component that is shared between space and environment and that jointly explains variation in species distributions (i.e. fraction [*b*] in Fig. 1) is the source of inflated type I error rates when assessing environment without controlling for spatial dependence (i.e. if only the component [*ab*] is tested). By testing fraction [*a*], we are removing this joint spatial component (i.e. $[a] = [ab] - [b]$) from the numerator of the *F*-statistic and hence exerting a control over the inflated test. A less acknowledged view is that the spatial 'legacy' in species distributions (i.e. fraction [*c*]) is also important (see Borcard *et al.*, 2004) and should be taken into consideration while challenging our ecological models. The most relevant question in this context is related to the origin of that fraction of the spatial structure in community organization: is it due to species interactions, dispersal and/or the effect of environmental variables that are missing from the model (i.e. were not considered), or perhaps a mix of all these components? Describing (e.g. plotting on maps) and attempting to understand the scales at which species and communities are organized can provide important clues about the origin of their spatial organization. Indeed, it is often the case that environmental variation is not sufficient to determine how patterns of species distributions are structured in space (e.g. Gravel *et al.*, 2008) and hence we need to further explore the data at hand. Due to advances in GIS techniques, we should routinely attempt to match unique spatial patterns (variation in fraction [*c*]) as a way to question the unmeasured processes (McIntire & Fajardo, 2009). One avenue, for instance, would be to consider connectivity measures (e.g. see Bender *et al.*, 2003, for a review of these measures) to assess the likelihood that dispersal is an important driver. Another possibility is the use of expert opinion (Low-Choy *et al.*, 2009) to help determine the likelihood of different unmeasured processes by analysing the spatial patterns of species distributions.

It is well accepted and demonstrated (and also shown here) that the presence of spatial structure in community and environment generates bias in statistical procedures (Bini *et al.*, 2009). However, there is some debate in the literature as to whether spatial correlation also affects parameter estimates (i.e. the strength of the association between community and environment). Hawkins *et al.* (2007,, and references therein), for instance, found that ordinary least squares (OLS) regression, which is used in RDA and CCA, is not seriously affected by spatial correlation in the sense that sample parameter estimates from spatially correlated populations are not biased. However, test procedures for predictor relevance (e.g. *t*-test for slopes, permutation procedures for the environmental component [*ab*]) are affected because standard errors of parameter estimates are smaller under spatial correlation than under spatial independence for the same number of degrees of freedom, hence generating smaller confidence intervals and therefore increasing type I error rates. Therefore, under spatial correlation, only standard errors are affected in OLS procedures but not slope estimates (Hawkins *et al.*, 2007). Since, it is generally recognized (e.g. Bini *et al.*, 2009) that OLS regression slopes change (shift) under non-spatial and spatial models, one conclusion would be to test slope significance under a spatial model but report slopes under a non-spatial model. In the case of variation partitioning, one could therefore test the importance of the link between communities and environment by testing fraction [*a*] (i.e. correcting for spatial autocorrelation), but the strength of the relationship should be reported on the basis of component [*ab*] (original estimate) and not fraction [*a*]. Whether parameter estimates may be affected by correlation depends on the model and type of statistic. Simulation work is still needed to determine the situations (types of model, ordination, degree and sign – positive or negative – of the spatial correlation) in which incorrect estimates are also generated. However, our point of view is that if indeed parameter estimates are not affected, the interest in reporting the spatially corrected estimate (i.e. fraction [*a*]) or not (component [*ab*]) may depend on what fraction [*b*] means. If spatial correlation in community data is partially or completely due to the measured environmental variables, which in turn are spatially structured, then perhaps reporting [*ab*] would be more appropriate, since fraction [*b*] would represent the spatialized component of the environment that induced correlation in the community in the first place. On the other

hand, if there are important missing predictors that are themselves spatially structured, then fraction [*b*] represents the covariation between the measured environment and unmeasured drivers and hence, as in any regression model in which coefficients are partial, fraction [*a*] should be then reported. However, determining which of these two possibilities is more likely is often not an easy task.

We hope that our simulation study and discussion of the issues revolving around spatial correlation and community analysis will provide an important instrument to guide ecologists in their decisions regarding the use of variation partitioning as a tool to explore patterns in community distribution. Our simulation protocols should also be useful while benchmarking other spatial methods in the context of variation partitioning or any other technique to estimate and control for spatial correlation in the study of ecological communities. The conclusions of the present research apply to time-series and phylogenetic data as well as spatial data.

ACKNOWLEDGEMENTS

Both authors would like to thank NSERC for funding support for this research. Discussions with Daniel Borcard, Stephane Dray and Dan Griffith were much appreciated. We would like to thank David Currie, Alexandre Diniz-Filho and three anonymous referees for their comments which greatly helped to improve this manuscript. We would like to dedicate this paper to the memory of biostatistician Nathan Mantel in the year of the 90th anniversary of his birth. Mantel, without knowing, has provided much fuel for the discussions surrounding techniques to describe and test for environmental and spatial variation in ecological communities.

REFERENCES

Anderson, M.J. & Gribble, N.A. (1998) Partitioning the variation among spatial, temporal and environmental components in a multivariate data set. *Australian Journal of Ecology*, **23**, 158–167.

Bender, D.J., Tischendorf, L. & Fahrig, L. (2003) Evaluation of patch isolation metrics for predicting animal movement in binary landscapes. *Landscape Ecology*, **18**, 17–39.

Bini, L.M., Diniz-Filho, J.A.F., Rangel, T.F.L.V.B. *et al.* (2009) Coefficient shifts in geographical ecology: an empirical evaluation of spatial and non-spatial regression. *Ecography*, **32**, 193–204.

Bivand, R. (1980) A Monte Carlo study of correlation coefficient estimation with spatially autocorrelated observations. *Quaestiones Geographicae*, **6**, 5–10.

Blanchet F.G., Legendre, P. & Borcard, D. (2008) Forward selection of explanatory variables. *Ecology*, **89**, 2623–2632.

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

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

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

ter Braak, C.J.F. (1986) Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, **67**, 1167–1179.

ter Braak, C.J.F. & Šmilauer, P. (2002) *Canoco reference manual and CanoDraw for Windows user's guide: software for canonical community ordination (Version 4.5)*. Microcomputer Power, Ithaca, NY.

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

Dale, M.R.T. & Fortin, M.-J. (2002) Spatial autocorrelation and statistical tests in ecology. *Ecoscience*, **9**, 162–167.

Deutsch, C.V. & Journel, A.G. (1992) *GSLIB: geostatistical software library and user's guide*. Oxford University Press, New York.

Diniz-Filho, J.A.F & Bini, L.M. (2005) Modelling geographical patterns in species richness using eigenvector-based spatial filters. *Global Ecology and Biogeography*, **14**, 177–185.

Dormann, C.F., McPherson, J., Araújo, M.B., Bivand, R., Bolliger, J., Carl, G., Davies, R.G., Hirzel, A., Jetz, W., Kissling, W.D., Kühn, I., Ohlemüller, R., Peres-Neto, P., Reineking, B., Schröder, B., Schurr, F.M. & Wilson, R. (2007) Methods to account for spatial autocorrelation in the analysis of distributional species data: a review. *Ecography*, **30**, 609–628.

Dray, S., Legendre, P. & Peres-Neto, P.R. (2006) Spatial modeling: a comprehensive framework for principal coordinate analysis of neighbor matrices (PCNM). *Ecological Modelling*, **196**, 483–493.

Dutilleul, P. (1993) Modifying the *t* test for assessing the correlation between two spatial processes. *Biometrics*, **49**, 305–314.

Fortin, M.-J. & Dale, M.R.T. (2005) *Spatial analysis. A guide for ecologists*. Cambridge University Press, Cambridge.

Gittins, R. (1985) *Canonical analysis – a review with applications in ecology*. Springer-Verlag, Berlin.

Gravel, D., Beaudet, M. & Messier, C. (2008) Partitioning the factors of spatial variation in regeneration density of shade-tolerant tree species. *Ecology*, **89**, 2879–2888.

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

Hawkins, B.A., Diniz-Filho, J.A.F., Bini, L.M., De Marco, P. & Blackburn, T.M. (2007) Red herrings revisited: spatial autocorrelation and parameter estimation in geographical ecology. *Ecography*, **30**, 375–384.

Keitt, T.H., Bjørnstad, O.N., Dixon, P.M. & Citron-Pousty, S. (2002) Accounting for spatial pattern when modeling organism–environment interactions. *Ecography*, **25**, 616–625.

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

Legendre, P. & Legendre, L. (1998) *Numerical ecology*, 2nd English edn. Elsevier Science BV, Amsterdam.

Legendre, P., Dale, M.R.T., Fortin, M.-J., Gurevitch, J., Hohn, M. & Myers, D. (2002) The consequences of spatial structure for the design and analysis of ecological field surveys. *Ecography*, **25**, 601–615.

Legendre, P., Borcard, D. & Peres-Neto, P.R. (2005) Analyzing beta diversity: partitioning the spatial variation of community composition data. *Ecological Monographs*, **75**, 435–450.

Legendre, P., Mi, X., Ren, H., Ma, K., Yu, M., Sun, I. F. & He, F. (2009) Partitioning beta diversity in a subtropical broad-leaved forest of China. *Ecology*, **90**, 663–674.

Lichstein, J.W., Simons, T.R., Shriner, S.A. & Franzreb, K.E. (2002) Spatial autocorrelation and autoregressive models in ecology. *Ecological Monographs*, **72**, 445–463.

Low-Choy, S., O'Leary, R. & Mengersen, K. (2009) Elicitation by design in ecology: using expert opinion to inform priors for Bayesian statistical models. *Ecology*, **90**, 265–277.

Martin, T.G., Wintle, B.A., Rhodes, J.R. Kuhnert, P.M. Field, S.A., Low-Choy, S.J., Tyre, A.J. & Possingham, H.P. (2005) Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecology Letters*, **8**, 1235–1246.

McIntire, E.J.B. & Fajardo, A. (2009) Beyond description: the active and effective way to infer processes from spatial patterns. *Ecology*, **90**, 46–56.

McQuarrie, A. & Tsai, C.-L. (1999) Model selection in orthogonal regression. *Statistics and Probability Letters*, **45**, 341–349.

Mundry, R. & Nunn, C.L. (2009) Stepwise model fitting and statistical inference: turning noise into signal pollution. *The American Naturalist*, **173**, 119–123.

Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Simpson, G.L., Solymos, P., Stevens, M.H.H. & Wagner, H. (2008) *Vegan: community ecology package*. R package version 1.15-0. Available at: http://cran.r-project.org/, http://vegan.r-forge.r-project.org/.

Olden, J.D., Jackson, D.A. & Peres-Neto, P.R. (2001) Spatial isolation and fish communities in drainage lakes. *Oecologia*, **127**, 572–585.

Peres-Neto, P.R., Legendre, P., Dray, S. & Borcard, D. (2006) Variation partitioning of species data matrices: estimation and comparison of fractions. *Ecology*, **87**, 2614–2625.

Rao, C.R. (1964) The use and interpretation of principal component analysis in applied research. *Sankhyaa (A)*, **26**, 329–358.

Rowe, R.J. & Lidgard, S. (2009) Elevational gradients and species richness: do methods change pattern perception? *Global Ecology and Biogeography*, **18**, 163–177.

Unwin, D. (1978) *An introduction to trend surface analysis*. Institute of British Geographers, Norwich.

Wartenberg, D. (1985) Multivariate spatial correlation: a method for exploratory geographical analysis. *Geographical Analysis*, **17**, 263–283.

Wilkinson, L., & Dallal, G.E. (1981) Tests of significance in forward selection regression with an F-to-enter stopping rule. *Technometrics*, **23**, 377–380.

Yamanaka, T., Tanaka, K., Hamasaki, K., Nakatani, Y., Iwasaki, N., Sprague, D.S. & Bjørnstad, O.N. (2009) Evaluating the relative importance of patch quality and connectivity in a damselfly metapopulation from a one-season survey. *Oikos*, **118**, 67–76.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Additional details on Moran's eigenvector maps (MEM) and related methods.

Appendix S2 Details on the forward selection procedure for orthogonal predictors in linear regression.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be reorganized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

BIOSKETCHES

Pedro R. Peres-Neto (PhD) is a community ecologist working on quantitative ecology with an interest in examining the roles of multiple ecological factors such as species-level traits (e.g. morphology, dispersal capacity, life history, phenotypic integration), habitat choice, landscape structure and species interactions in driving species distributions and community structure.

Pierre Legendre (PhD) is interested in understanding the processes that determine the spatial structure of natural communities and control beta diversity. He has developed a number of statistical methods to answer questions that arise in this research framework.

Editor: José Alexandre F. Diniz-Filho