# A UNIFIED STRATEGY FOR ESTIMATING AND CONTROLLING SPATIAL, TEMPORAL AND PHYLOGENETIC AUTOCORRELATION IN ECOLOGICAL MODELS

Pedro R. Peres-Neto

Université du Québec à Montréal. Case postale 8888, succursale Centre-ville. Département des sciences biologiques, Faculté des sciences. Montréal (Québec) H3C 3P8. Canada.

peres-neto.pedro@uqam.ca

## ABSTRACT

The goals of this paper are to expose ecologists to the problem related to statistical inference when testing the association between data sets that are autocorrelated and to introduce a relatively new method for controlling the bias introduced by autocorrelation that can be easily incorporated in any statistical approach. In addition, I show the flexibility of this class of methods to the types of data that ecologists are currently most interested, namely temporal, spatial and phylogenetic data. In this contribution, I also stress the point that is not all variation due to autocorrelation that affects statistical inference and is important to control only the component that biases inference. Thus, statistical frameworks should attempt to separate the autocorrelation component that biases inference from the one that may prove interesting for understanding important ecological processes, such as contagious processes, driving spatial patterns in species distributions.

**Key words:** Statistical inference, Predictors, Autocorrelation, Eigenfunction analysis.

## RESUMO

**UMA ESTRATÉGIA UNIFICADA PARA A ESTIMATIVA DE COMPONENTES ESPACIAIS, TEMPORAIS E FILOGENÉTICOS EM MODELOS ECOLÓGICOS.** O objetivo deste trabalho é de expor aos ecólogos o problema relacionado aos testes de inferência estatística quando os dados são autocorrelacionados e apresentar uma técnica relativamente nova que pode ser facilmente incorporada em análises estatísticas para controlar os erros causados pela autocorrelação. Além disso, eu demonstro a flexibilidade deste método utilizando três tipos de dados que são importantes em análises ecológicas: dados temporais, espaciais e filogenéticos. Neste trabalho, eu reitero que não é toda a variação autocorrelacionada que afeta as inferências estatísticas e que é importante controlar apenas o componente de variação reponsável. Assim, análises estatísticas devem ser realizadas com o objetivo de separar o componente de variação autocorrelacionada – que causa erros em testes de hipóteses – do componente que pode ser importante para a compreensão de processos ecológicos, como processos contagiosos (e.g., dispersão), estruturando padrões de distribuição espacial em espécies.

**Palavras-chaves:** Inferência estatística, Preditores, Autocorrelação, Análise eigenfunction.

## INTRODUCTION

Testing and estimating the level of association between two or more variables or two or more multivariate data sets is a long-standing approach in identifying important processes governing evolutionary and ecological patterns. For instance, community ecologists seek to establish relationships between environmental characteristics and species distribution (e.g., Jackson & Harvey 1993, Rodríguez & Lewis 1997, Jenkins & Buikema 1998). Ecomorphologists often test if size and shape variation are correlated to ecological differences among species (e.g., Losos 1990, Douglas & Matthews 1992, Van Damme *et al.* 1998). Among systematists a common goal is to determine whether or not spatial distribution is related to phenotypic or genetic differentiation among populations or species (e.g., Douglas & Endler 1982, Douglas *et al.* 1999). These research programs embrace rather different questions and types of multivariate data, but they all

involve comparisons between two or more variables or data sets in order to measure their degree of association. If statistically significant, the match between data sets contributes to evidence about the processes determining the association.

Perhaps the most stringent assumption in any parametric, non-parametric and distribution-free methods (see Peres-Neto & Olden 2001, for a distinction between these type of methods) used to estimate the association between data sets (e.g., correlation, multiple regression, redundancy analyses, canonical correspondence analysis, canonical correlation analysis, Mantel test, Procrustean rotations, to mention a few) is the independence of sampled observations (i.e., each observation must be identically independent of each other). Parametric (e.g., t-distribution, F-distribution) and non-parametric (e.g., Mann-Whitney, rank correlations) sampling distributions were developed by assuming that every observation in the sample was drawn randomly and independently from each other. In permutation tests, the assumption of independence is also relevant (see Manly 1997) as observations are randomly permuted in relation to each other when testing, for example, the significance of correlation or regression slopes.

Independence of observations entails that no observation in a sample can be predicted by another observation in the same sample and that the best predictor of any observation is simply the mean. When observations are not independently sampled from each other, they are said to be autocorrelated and in this case an observation can be predicted as a function of other observations. For instance, consider a random variable under spatial dependence (e.g., the abundance distribution of a particular species). Two pairs of observations may have values that are more similar (positive autocorrelation) or less similar (negative autocorrelation) according to their geographic distance to a greater extent than one would expect if the differences among those values were due to chance alone or to other predictors of interest (e.g., environment). Therefore, in the presence of autocorrelation, the number of degrees of freedom in the sample is smaller than when observations are independent. As a consequence, association tests generate unrealistic significance estimates because a larger number of degrees of freedom than the appropriate one is used, thus generating narrow confidence limits for hypothesis testing, making the test less conservative than expected by pure chance when the null hypothesis is true. In order words, the nominal type I error is greater than the pre-established one (i.e., significance level or alpha). In addition, autocorrelation can also promote bias in estimates (e.g., slope). We will take a closer look into this issue in the next section where the problem is demonstrated by means of simulation.

Autocorrelation is common in nature and operates either as a factor moulding or constraining ecological variables, or as a confounding variable that introduces bias by influencing the interpretation of statistical models (Clifford *et al.* 1988, Dutilleul 1993). In Ecology, three important processes that may cause autocorrelation have drawn a great deal of attention in recent years: spatial, temporal and phylogenetic variation (Ives & Zhu 2006). Although these processes can introduce bias in ecological models, they can be also interesting on their own. For instance, geographically contagious biotic processes such as dispersal may promote spatial autocorrelation in species distributions that may cause bias in models, but they are also interesting as an ecological process (e.g., the study of ecological factors driving dispersal differences among species). The analysis of time series (see Bence 1995) of population abundances may also cause problems due to autocorrelation caused, for instance, by density-dependent processes (i.e., abundances tend to be correlated through time), but it also allows ecologists to understand, for instance, population dynamics and predict the fate of populations. In the case of phylogenetic analysis, controlling for autocorrelation may allow us to study correlated evolution due to selective processes but it can be also used to assess the level of plasticity or canalization of ecological features (e.g., feeding mode, morphology, behavior) across species.

Given the great deal of attention that was given to the problems caused by autocorrelation, it seems that ecologists do not recall at times that autocorrelation in itself may be caused by interesting ecological phenomena (e.g., contagious processes, density dependence, evolution) and we should not perceive autocorrelation always as a problem. In this paper, I advocate a balanced view where not all the variation in ecological data that is autocorrelated should be

eliminated in order to allow for unbiased hypothesis testing. Only the component of variation due to autocorrelation that causes bias in statistical tests should be removed, whereas the component that does not should be left in the data and further analyzed to address interesting questions regarding the processes driving this variation. Note, however, that the amount of autocorrelation that may bias statistical testing will vary among data sets and it is not always possible to eliminate the component that causes bias and still have some autocorrelated variation left for interpretation. There are numerous techniques that aim at controlling the effects of autocorrelation in ecological models (Legendre 1993, Diniz-Filho 2000, Dale & Fortin 2002, Martins *et al.* 2002), but most of them are designed only for controlling autocorrelation. The goal of this paper is two-fold: (1) show how the problems in hypothesis testing arise under autocorrelation, and (2) describe a statistical approach that attempts to control for the autocorrelation component of ecological variation that affects hypothesis testing. The method is flexible enough to tackle three of the most important ecological processes that may generate dependence among observations, namely spatial, temporal and phylogenetic autocorrelation, providing a unified strategy for estimating and controlling autocorrelation in ecological models. Moreover, the class of method presented here is flexible enough that it can be applied to any type of distribution under generalized linear model procedures (e.g., analysis of variance, logistic/ binomial and Poisson regressions) and also applicable to non-parametric modeling tools such as regression and classification trees (CART) and artificial neural networks (see Elith *et al.* 2006 for a review of novel modeling techniques in ecology). Examples dealing with spatial, temporal and phylogenetic variation are provided.

## HOW DOES AUTOCORRELATION AFFECT ECOLOGICAL MODELS?

### *THE ANALYSIS OF TIME SERIES*

Let us suppose that an ecologist is interested in testing whether the abundance of a lake fish species is controlled by total zooplankton abundance based on a time series of 10 years with samples collected every 2 months in a lake, totalizing 60 observations. To test for their association, a regression slope between fish and zooplankton abundance will be used based on an alpha level of 0.05. If regression residuals are independent and normally distributed, and if the population slope (i.e., true slope between fish and zooplankton in the study lake) is zero, there is a 5 % (i.e., alpha = 0.05) chance that a sample slope of 60 observations randomly selected through time will be significant even though the population value is zero. This chance of committing a type I error is known and established *a priori* (i.e., alpha). However, if fish and zooplankton are temporally autocorrelated, the chance of sampling 60 observations and finding a significant slope between them is greater than 0.05 (in some cases much greater, e.g., 0.30).

In order to show the problem in a more compelling manner, I will use a simulation to demonstrate the problem. Assume that the abundances of the fish species and zooplankton are independent (slope = 0) of each other but that their abundances are regulated by intra-taxa density dependence under a stochastic logistic model (or a first-order nonlinear autoregression model; see Dennis & Taper 1994, Clark & Bjørnstad 2004) as follows:

$$\ln(N_t) = \ln(N_{t-1}) + b_0 + b_1 \exp(\ln(N_{t-1})) + ze$$

where $b_0$, $b_1$ and $z$ are constants, and $e$ is a normally distributed N(0,1) random shock (i.e., mean=0 and variance=1) to the population growth rate (e.g., environmental noise).

Now, let us simulate a time series for the fish species fish and total zooplankton abundance as follows:

$$\ln(N_t fish) = \ln(N_{t-1} fish) + 0{,}5 - 0.01\exp(\ln(N_{t-1} fish)) + 0.1e$$

$$\ln(N_t zoo) = \ln(N_{t-1} zoo) + 0{,}5 - 0.01\exp(\ln(N_{t-1} zoo)) + 0.3e$$

The initial abundances were set as $N_0 fish = \ln(70)$ and $N_0 zoo = \ln(1000)$. Two simulated time series using these equations are shown in Fig. 1. Note that if another time series is generated either for fish or zooplankton, the temporal trajectories would be different and independently generated from the one presented in Fig. 1 due to the random shock introduced by *e*. Because regression slopes of the simulated time series represent independent realizations of the same statistical

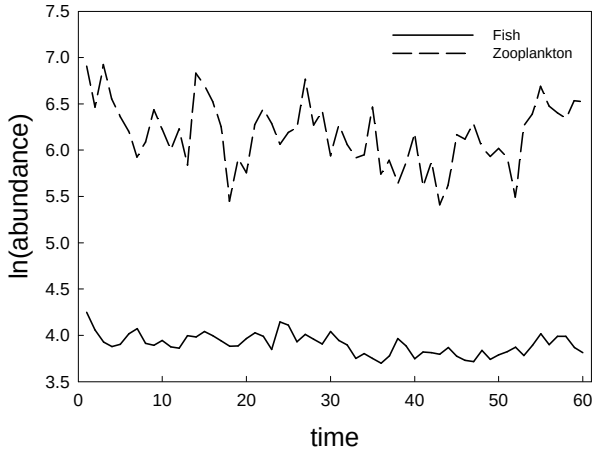



Figure 1. Two simulated time series using a first-order nonlinear autoregression model to generate density dependence in fish and zooplankton. Models are described in the text.

population, they can be used to illustrate the bias in statistical tests for regression slopes. In this case, because zooplankton and fish time series are independently generated, the population slope is, by definition, zero. I have generated 10,000 time series based on the above formulae and tested the slope based on the parametric *t* test and a permutation test (999 permutations), as presented by Peres-Neto & Olden (2001). The frequency of significant tests based on an alpha value of 0.05 was 0.1199 (1,199 rejections/10,000 time series) for the *t* test and 0.1280 (1,280 rejections/ 10,000 time series) for the permutation test. These frequencies represent type I error estimates, i.e., the sampling frequency at which the null hypothesis will be rejected when it is true, given that the slope in the population is zero. The significance level, or alpha value, established *a priori* is the probability of committing the so-called type I error. In other words, if a significance level of 0.05 is chosen, for 10,000 sample values of the test being conducted, 500 of them will be considered significant when in reality (i.e., in the population) they are not. If we had used the following non-density dependent processes

$$\ln(N_t fish) = \ln(70) + 0.5 - 0.01\exp(\ln(70)) + 0.1e$$

$$\ln(N_t zoo) = \ln(1000) + 0.5 - 0.001\exp(\ln(1000)) + 0.3e$$

for generating fish and zooplankton abundances, respectively, instead, the frequency of significant slope tests based on an alpha value of 0.05 was 0.0499 (499 rejections/10 000 time series) for the *t* test and 0.051 (510 rejections/ 10 000 time series) for the permutation test. Thus, simply because the data points were not independently generated in the first set of simulations, the sampling tests tended to reject more often than the pre-established significance level (alpha). Hence, because of the autocorrelation between the data points, statistical tests become biased as they present greater nominal type I error rates than the pre-established significance level. Under extreme circumstances, the estimation of the slope may be also biased.

*THE ANALYSIS OF SPATIAL DATA*

Perhaps one of the most innate patterns in natural systems is their spatial organization. The issues of autocorrelation in modelling ecological data have been of particular interest when considering spatial distributions given that one of the most common routines in ecological investigations is to collect data on species distribution and environmental data across space to assess how habitat features drive species distribution across a particular landscape of interest. The problem of autocorrelation in this case is manifested by the fact that spatial processes may influence both species' distributions and environmental factors generating apparent species-environment concordance (Legendre 1993). Species abundances are spatially organized across landscapes due to ecological contagious processes such as population growth, geographic dispersal, differential fertility or mortality, social organization or competition dynamics, for instance. Environmental factors are also often spatially organized across landscapes where nearby sites tend to contain more similar habitat conditions than distant ones. Establishing relationships between species distributions and environmental characteristics is a widely-used approach (e.g., Legendre & Fortin 1989, Jackson & Harvey 1993, Diniz-Filho & Bini 1996, Rodríguez & Lewis 1997) in the search for causes dictating patterns in species distributions. Habitat models relating habitat characteristics and community structure (species occurrence or abundance) are expected to answer at least two questions: (1) How well is the distribution of a set of species explained by the given set of predictive variables? and (2) Which variables are irrelevant or redundant in the sense of

failing to strengthen the explanation of patterns after certain other variables have been taken into account? The first question relates to the predictive power of the model that can be used in conservation management, for questions such as estimating habitat suitability, forecasting the effects of habitat change due to human interference, establishing potential locations for species re-introduction, or predicting how community structure may be affected by the invasion of exotic species. The second question is important for heuristic issues such as determining the likelihood of competing hypotheses to explain particular patterns in community structure (Peres-Neto *et al.* 2001).

Regardless of the goal, both questions involve statistical tests that may prove challenging under spatial autocorrelation. In order to demonstrate the problem of testing the relationship between spatially autocorrelated processes, I have applied a very simple method for generating spatially dependent data (but see Legendre *et al.* 2005 for other possibilities). First, I generated a matrix **Y** containing two normally distributed variables N (0.1) with ten observations each. Then, for each observation in **Y**, I have generated nine other observations that were created by adding small normally distributed deviates –N (0.1)/15 to each of the ten original observations; after that, matrix **Y** contained two variables and 100 observations. A second matrix **X** was generated in the same manner. The spatial patterns depicted by two independent realizations of this process (i.e., matrices **X** and **Y**) are shown in Fig. 2. Note that we can distinguish well the ten clusters of ten observations each generated by the simulation process in both matrices. Here, because the goal is simply to show the inferential problems under spatial autocorrelation, I did not generate abundance-like data and I will simply test the association between matrices **Y** and **X**.

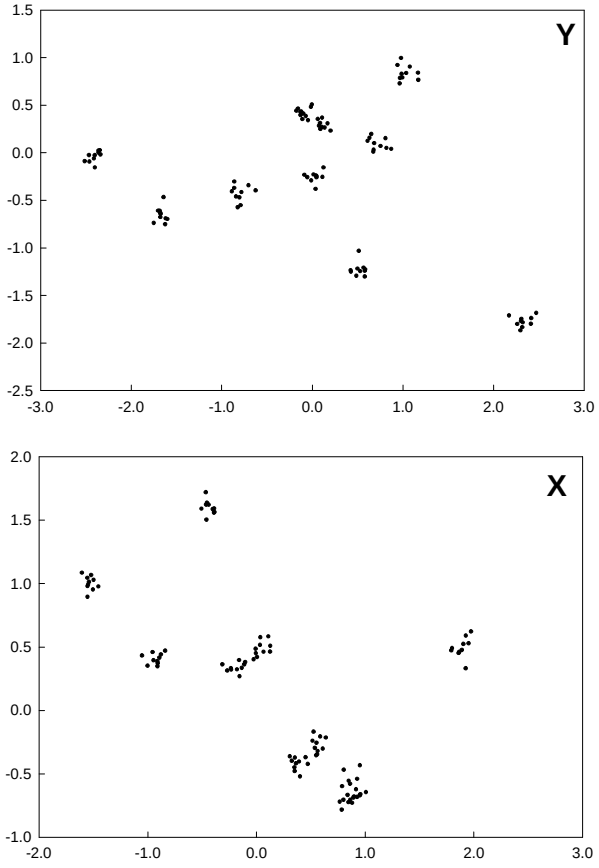



Figure 2. Spatial patterns depicted from two independent realizations of a simple process to generate spatial autocorrelation (see text for a description of the process). This process was used to generate two sets of matrices (**X** and **Y**) to test for their correlation when data matrices are spatially autocorrelated. Axes represent artificial geographic coordinates.

Perhaps canonical analyses such as redundancy analysis (RDA, Rao 1964), canonical correspondence analysis (CCA, ter Braak 1986), and distance-based redundancy analysis (db-RDA, Legendre & Anderson 1999) are the most commonly used tool for modeling communities through environmental predictors. Canonical analyses can be best understood as methods for extending multiple regression, which has a single response **y** and multiple predictors **X** (e.g., several environmental predictors), to multiple regression involving multiple response variables **Y** (e.g., several species) and a common matrix of predictors **X**. I have applied here a RDA to test the association between matrices **Y** and **X** simulated above. I have generated 10,000 **Y** and **X** matrices based on the process described above for generating spatially autocorrelated data and tested their association using the redundancy statistic $R^2_{Y|X}$ based on a well-established permutation test (999 permutations were used) described elsewhere (Manly 1997, Legendre & Legendre 1998, Peres-Neto *et al.* 2006). As for the time series simulation, because **Y** and **X** represent independent realizations, they can be used to illustrate the bias of statistical tests for under autocorrelation. Because **Y** and **X** are independently generated, the population is also, by definition, zero. The frequency of significant tests based on an alpha value of 0.05 was 0.9431 (9431 rejections/10,000 time series), indicating an extremely high nominal type I error

rate. On the other hand, the test for based on **Y** and **X** matrices containing only two normally distributed variables N(0,1) with 100 observations each (i.e., without the small deviates around the original observations) provided a correct type I error (0.0498, 498/10,000 tests).

*THE ANALYSIS OF PHYLOGENETIC DATA*

Ecologists and evolutionary biologists are commonly interested in testing the association between variables where the observations represent species (usually mean values for each species). Examples of these type of data include the relationship between gestation time and time to sexual maturity in primates (Harvey *et al.* 1987), the relationship between home range and body size (Garland *et al.* 1992) and the relationship between behavioural and morphological characters (Losos 1990), to mention just a few. However, because closely related species tend to be more similar to each other due to evolutionary processes, they may not represent independent observations that can be directly used to assess these relationships (Felsenstein 1985). In this case, variables of interest (e.g., body size) are said to be phylogenetically autocorrelated.

There are different ways of generating phylogenetically autocorrelated data (Martins & Garland 1991, Fleckleton *et al.* 2002, Martins *et al.* 2002) and they are used to assess the performance of comparative methods in correcting the problem of inflated type I errors due to autocorrelation. Most methods assume a model of evolution such as Brownian movement (BM) or Ornstein-Uhlenbeck (OU) process (Felsestein 1988, see also Diniz-Filho 2000 for a discussion on these processes) that generates data under independence evolution by assuming that characters evolve by drift (i.e., without selection in the case of BM) or different levels of constraints such as stabilizing selection towards an optimum across phylogenetic lineages (in the case of OU). Therefore, one of the goals of testing the correlation between characters under a comparative framework is to test whether the correlation between characters is beyond what is expected by independence evolution and hence estimate the importance of correlated evolution due to selection in driving these correlations. Regardless of the model used, the evolution of a character is proportional to the shared evolutionary history due to common ancestry. In the case of Brownian movement, the evolution is directly proportional to the shared history, whereas in the OU process the common history is lessened.

There are different computational implementations in order to generate phylogenetically autocorrelated data under the BM and OU processes. In order to illustrate the problem, I suggest yet another implementation that generates data similar to a BM process. I will not demonstrate that the algebra behind this new implementation generates data akin to a BM process, but it follows the work developed by Garland & Ives (2000), Butler *et al.* (2000) and Rohlf (2001). Assume the phylogenetic tree depicted in Fig. 3a and its associated phylogenetic covariance matrix (Fig. 3b), which is calculated directly from the tree. The main diagonal of the covariance matrix represents the variance, which is calculated as the distance from the root to tip (i.e., the total time of evolution). In this case, the variance is 14.3. The covariance values (off diagonal) are the shared evolution between any given two species. For instance, *Anolis opalinus* and *Anolis grahami* share 12.2 path lengths (arbitrary units of time; 1.0 + 6.8 + 0.5 + 2.0 + 1.9 = 12.2). Phylogenetically autocorrelated data for two independent characters **x** and **y** were generated as follows:

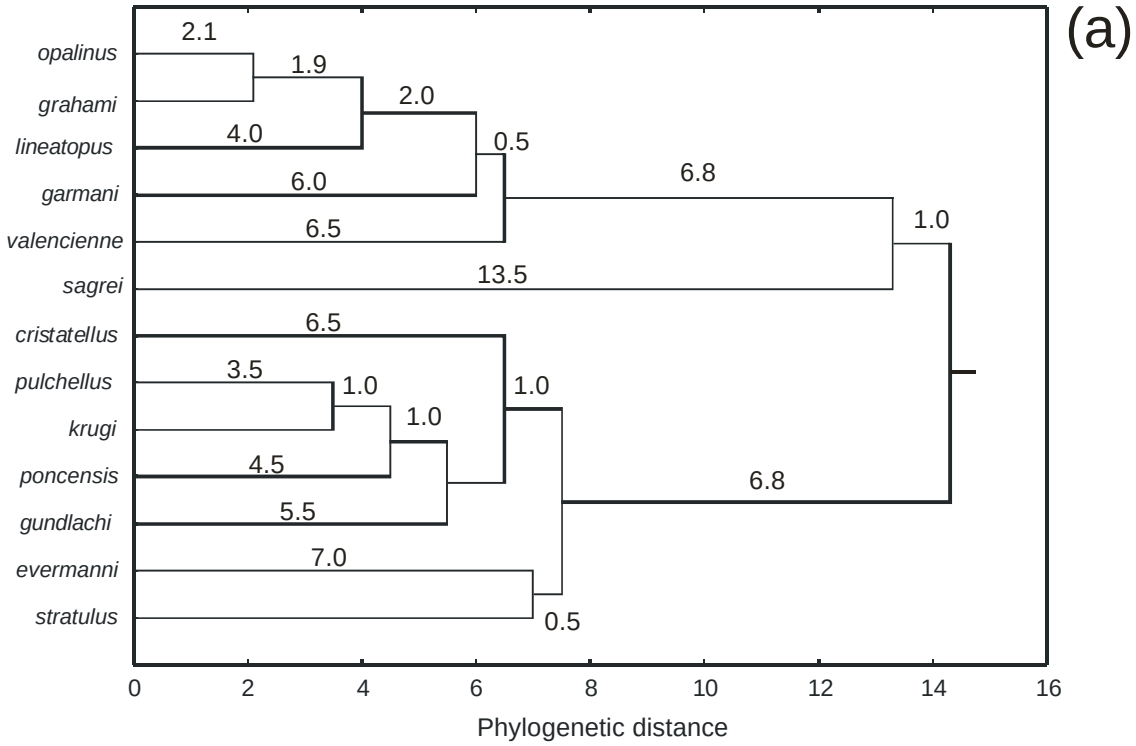
$$\mathbf{x} = \mathbf{e}\mathbf{V}^{1/2} \text{ and } \mathbf{y} = \mathbf{e}\mathbf{V}^{1/2}$$

where **e** is a (1,n) vector containing n (number of species) normally distributed variables N(0,1) and $\mathbf{V}^{1/2}$ is the root (here I used a Cholesky decomposition) of the covariance matrix.

I have generated 10,000 **x** and **y** vectors sets and the rate of rejection of a t-test for their correlation based on an alpha equal to 0.05 was 0.1655 (1655 rejections/ 10 000 character sets). A permutation test (Peres-Neto & Olden 2001) provided a rate of 0.1574. Note that **x** and **y** were independently generated, though the rejection rate of their correlation test was much higher than the expected 0.05. Next, I simulated data sets where only **x** is phylogenetically autocorrelated and **y** is simply a vector of random normally distributed N(0,1) values (i.e., without autocorrelation). In this case, the rejection rate was 0.049 (i.e., 490 rejections/10 000 character sets) indicating that only when both variables are autocorrelated that statistical inference is biased.

The fact that both variables or data sets involved in the comparison need to be autocorrelated to affect hypothesis testing is not unique to phylogenetic data and this fact has been already stressed in the literature in the case of spatial data (e.g., Dutilleul 1993, Legendre *et al.* 2004).

I hope to have convinced the readership of the problems related to statistical testing of autocorrelated data and that this type of data has to be properly analyzed. In the next section, I introduce a method for filtering out (removing) the autocorrelated variation that causes bias in statistical testing.




(b)

| | *opalinus* | *grahami* | *lineatopus* | *garmani* | *valencienne* | *sagrei* | *cristatellus* | *pulchellus* | *krugi* | *poncensis* | *gundlachi* | *evermanni* | *stratulus* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *opalinus* | 14.3 | 12.2 | 10.3 | 8.3 | 7.8 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| *grahami* | 12.2 | 14.3 | 10.3 | 8.3 | 7.8 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| *lineatopus* | 10.3 | 10.3 | 14.3 | 8.3 | 7.8 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| *garmani* | 8.3 | 8.3 | 8.3 | 14.3 | 7.8 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| *valencienne* | 7.8 | 7.8 | 7.8 | 7.8 | 14.3 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| *sagrei* | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 14.3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| *cristatellus* | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 14.3 | 10.8 | 9.8 | 8.8 | 7.8 | 6.8 | 6.8 |
| *pulchellus* | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 10.8 | 14.3 | 9.8 | 8.8 | 7.8 | 6.8 | 6.8 |
| *krugi* | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 9.8 | 9.8 | 14.3 | 8.8 | 7.8 | 6.8 | 6.8 |
| *poncensis* | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 8.8 | 8.8 | 8.8 | 14.3 | 7.8 | 6.8 | 6.8 |
| *gundlachi* | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 7.8 | 7.8 | 7.8 | 7.8 | 14.3 | 6.8 | 6.8 |
| *evermanni* | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 6.8 | 6.8 | 6.8 | 6.8 | 6.8 | 14.3 | 7.3 |
| *stratulus* | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 6.8 | 6.8 | 6.8 | 6.8 | 6.8 | 7.3 | 14.3 |

Figure 3. (a) Phylogenetic tree of 13 species of West Indian Anolis lizards (adapted from Losos 1990). Numerical values on the tree indicate branch length that are used to calculate the phylogenetic covariance matrix; (b) phylogenetic covariance matrix based on the Anolis phylogeny. Note that variance is 14.3 (main diagonal) and that covariance values (off diagonal) are the shared evolution between any given two species. For instance, Anolis opalinus and Anolis grahami share 12.2 path lengths (arbitrary units of time; 1.0 + 6.8 + 0.5 + 2.0 + 1.9 = 12.2).

## EIGENFUNCTION BASED FILTERING METHOD

The eigenvector filtering method has been suggested independently by different researchers both in the context of spatially (Griffith 2000, Borcard & Legendre 2002, Griffith & Peres-Neto 2006 reviewed this method in the case of spatial data) and phylogenetically autocorrelated data (Diniz-Filho *et al.* 1998). These implementations were distinct from the one presented here though. The filtering method begins with an eigenfunction decomposition of a truncated matrix (see below) representing the proximity among observations (i.e., temporal, spatial and species). These eigenvectors, corresponding to positive eigenvalues, are then used as spatial descriptors in regression or any type of analysis aiming at testing the association of autocorrelated data. In the original method described by Borcard & Legendre (2002) and Diniz-Filho *et al.* (1998), the truncated matrix of temporal, spatial or phylogenetic distances was built in such a way that it considered the influence of an observation on itself (e.g., the geographic and phylogenetic distance matrix has non-zero values in the main diagonal). Although this consideration may be seen as difficult to justify, there are examples of spatial models where it has been applied (Bavaud 1998). Here I use the implementation provided by Dray *et al.* (2006) where this problem is solved. The eigenvector procedure (after Dray *et al.* 2006) may be summarized with the following steps:

**1.** Compute a pairwise Euclidean distance matrix among observations units ($\mathbf{D} = \left[ d_{ij} \right]$);

**2.** For spatial data, choose a threshold value *t* and construct a truncated connectivity matrix **W** (i.e., not all observations are connected) using the following rule:

$$W = \left[ w_{ij} \right] = \begin{cases} 0 \text{ if } i = j \\ 0 \text{ if } d_{ij} > t \\ \left[ 1 - \left( d_{ij} / 4t \right)^2 \right] \text{ if } d_{ij} \leq t \end{cases}$$

where *t* is chosen as the maximum distance that maintains all sampling units being connected using a minimum spanning tree algorithm (Legendre & Legendre 1998). Other methods (see Dray *et al.* 2006 and Griffith & Peres-Neto 2006) are also available for establishing thresholds and different options can be explored to the specific problem at hand. However, simulation work (Peres-Neto & Legendre, unpublished data) indicates that the method based on minimum spanning tree corrects for the problem of inflated type I errors in spatially autocorrelated data.

In the case of temporal data that has been sampled according to a constant time interval, the procedure is simplified as **W** becomes simply a matrix where the distance between adjacent time periods is equal to 1 and the distance between non-adjacent periods is equal to 0 (i.e, $d_{ij} = 1$ only if i-j = -1 or 1, if i-j takes any other value, e.g., I = 3 and j = 1, then $d_{ij} = 0$). Note, however, that irregular sampling time periods can easily be incorporated by using the method described for spatial data where a pairwise distance matrix representing the difference between time periods is applied. In the case of phylogenetic data, matrix **W** can be easily calculated by subtracting the phylogenetic variance by each element in the variance-covariance matrix.

**3.** Compute the eigenvectors of the centered **W** matrix:

$$\mathbf{(I - 11^T/n)W(I - 11^T/n)}$$

The eigenvector matrix is a square matrix (i.e., a matrix with equal numbers of rows and columns) where the columns contain variables representing distinct temporal, spatial or phylogenetic patterns, depending on the application, and the rows represent observations (i.e., temporal samples, spatial sample or species). Given the non-Euclidean nature of **W**, both positive and negative eigenvalues are produced. The non-Euclidean part is introduced by the fact that only certain connections among observations, and not all, are considered in **W.** The extracted eigenvectors represent the decomposition of the Moran's index of autocorrelation (MI; see Legendre & Legendre 1998, for the use of MI in spatial data, and Diniz-Filho 2001 in phylogenetic data) into all mutually orthogonal and uncorrelated temporal, spatial or phylogenetic patterns. Eigenvectors having associated eigenvalues that are positive represent positive autocorrelation, whereas eigenvectors having negative eigenvalues represent negative autocorrelation. A MI for any eigenvector *v* can be directly calculated as follows:

$$MC(v) = \frac{n}{1^T S 1} v^T (I - 11^T/n) W (I - 11^T/n) v = \frac{n}{1^T S 1} v^T W v$$

In the context of spatial data, eigenvectors with large eigenvalues represent coarse scales of variability or landscape-wide trends (e.g., global); eigenvectors with intermediate size eigenvalues represent medium scales (e.g., regional); eigenvectors with small eigenvalues represent fine scales or patchiness (e.g., local). Therefore, the extracted eigenvectors capture a range of geographic scales encapsulated in a given dataset, restricted by the landscape boundary extent of sample locations and the threshold value used to truncate distance. The same analogy can be made to temporal and phylogenetic data. In the latter case, eigenvectors with large eigenvalues represent early speciation events, whereas eigenvectors with small eigenvalues represent later events. In order to provide a picture of the types of patterns that these eigenvectors represent, I have plotted the temporal patterns depicted by three selected eigenvectors constructed for a time series containing 100 observations (Fig. 4, top three panels).

The resulting eigenvectors, themselves, are then used directly as synthetic explanatory variables in the analysis. This modeling approach is semiparametric in

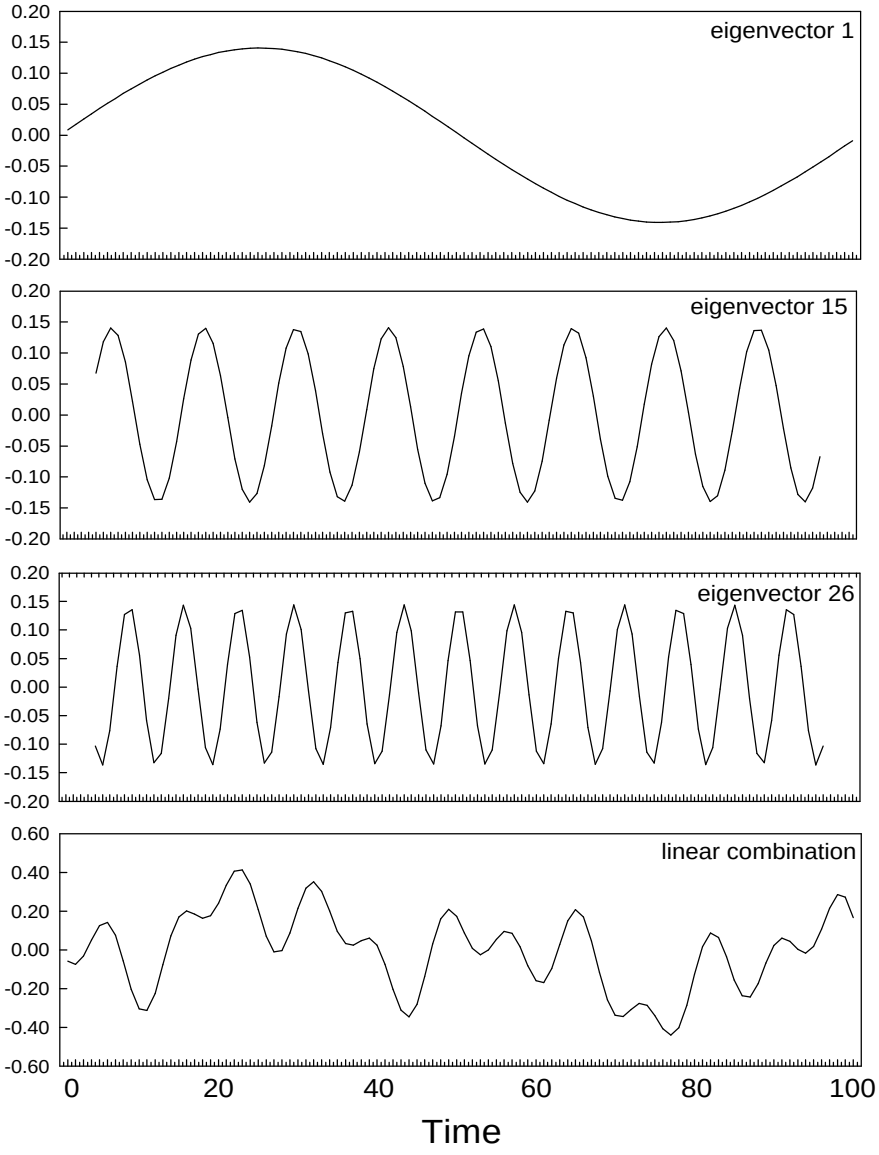



Figure 4. Temporal patterns depicted by three selected eigenvectors (1, 15 and 26) constructed for a time series containing 100 observations and by the sum of four randomly selected eigenvectors for the same time series. The latter series mimics a possible outcome of a linear combination of eigenvector maps created by a model selection technique in regression analysis.

nature, casting autocorrelation as some unknown function (nonparametric) - which must be estimated from a given dataset - that is additively coupled with a set of covariates whose coefficients need to be estimated (parametric). For instance, in a multiple regression model of **y** on a set of predictors, we add the eigenvectors as additional predictors into the model and linear combinations between them constitute an estimate of the unknown autocorrelation function. In practice (see next section), we only apply a judiciously selected subset of eigenvectors, since some of them may not estimate well the autocorrelation function of a particular data set. Here, and in other ecological applications, the set of candidate eigenvectors to be selected represents positive spatial autocorrelation (i.e., only eigenvectors having positive eigenvalues are retained for further analysis) which is the one known to inflate type I error rates (Legendre *et al.* 2004. Linear combinations of selected eigenvectors are capable of representing complex patterns in data. Fig. 4 (bottom panel) shows the sum of four randomly selected eigenvectors for the time series containing 100 observations, which mimics a possible outcome of a linear combination of eigenvector maps created with a model-selection technique in regression analysis. Note how the combination of eigenvectors can depict complex patterns of variation. Dray *et al.* (2006) and Griffith & Peres-Neto (2006) proposed model selection techniques for eigenvectors in the context of spatial data, but they are directly applicable to temporal and phylogenetic data. The selection procedure introduced by Griffith & Peres-Neto (2006) reduces the level of autocorrelation in regression residuals so that the assumption regarding independence is met, whereas the method proposed by Dray *et al.* (2006) selects eigenvectors based on a forward selection procedure aiming at maximizing the amount of autocorrelation explained by the eigenvectors. In the next section, I provide a complete example of application for temporal, spatial and phylogenetic data.

## APPLICATIONS OF EIGENFUNCTION FILTERING METHOD

### *ANALYSIS OF A TIME SERIES*

As an example, I analyzed a time series of yellow perch and chlorophyll concentration from Sparkling lake in Wisconsin, USA. Data comes from to the North Temperate Lakes – Long-term ecological research project (http://lter/limnology/wisc.edu) and details on these particular data and sampling procedures can be found in Beisner *et al.* (2003). Fig. 5a shows the temporal trends for yellow perch and chlorophyll. Although weak (Fig. 5b), there is a significant positive correlation between chlorophyll and yellow perch (a zooplanktivorous species). This suggests a top-down trophic cascade in which fish predation controls zooplankton, thereby reducing grazing pressure on phytoplankton, leading to an increase of chlorophyll concentration when perch is relatively more abundant. First, I tested for each variable separately whether all positive eigenvectors significantly explained its variation using a multiple regression model (perch: $F = 18.54$, $P = 0.0001$; chlorophyll: $F = 3.019$, $P = 0.0004$), indicating that both variables have a strong degree of temporal autocorrelation. In this case, all eigenvectors were used because selection procedures (e.g., forward selection) tend to inflate the overall significance of the model. Once an autocorrelation component is considered to be present in both variables, a selection procedure should be used. Next, I performed a forward selection for a multiple regression of chlorophyll on the positively autocorrelated eigenvectors, and 11 eigenvectors were found to be significant. Finally, in order to test for the effect of perch on chlorophyll, while controlling for temporal autocorrelation, I performed a multiple regression of chlorophyll on perch, and the 11 selected eigenvectors. The perch contribution was no longer significant to the model (slope significance = 0.9137), indicating that the initial conjecture of top-down control was due to temporal autocorrelation inherent to perch and chlorophyll dynamics.

### *ANALYSIS OF A SPATIAL DATA*

Here I analyze a fish data comprising the distribution of 27 species and environmental data in 53 sample sites of the river Macacu, Brazil (details on sampling procedures, species and environmental data are provided in Peres-Neto 2004). My goal here is to test whether species distributions are driven by the environmental variation found in the system. In this data, 12 eigenvectors having positive eigenvalues were

extracted and used as spatial descriptors. The eigenvectors were calculated on the basis of geographic distances between sites, though one could also consider the water-course distance within the river network, which can better represent spatial relationships in riverine systems (Olden *et al.* 2001).

In order to assess the significance of environmental predictors on species distributions, I used variation partitioning for redundancy analysis (RDA) as a template. An initial test of the significance of the RDA statistic $R^2_{Y|X}$ indicates that environment is a significant driver of species distribution ( $R^2_{(Y|X)adj}$ = 0.191, P = 0.001). Here, I report adjusted values () for the RDA statistic, as unadjusted values are highly biased (Peres-Neto *et al.* 2006). The adjusted RDA statistic parallels the adjusted $R^2$ (coefficient of determination) in linear regression analysis. First, I tested the significance of based on all 12 spatial eigenvectors which indicated a significant amount of autocorrelation in both species distribution and environmental matrices (species distribution: = 0.199, P = 0.001; environment: = 0.599, P = 0.001). Next, to filter out the influence of the spatial autocorrelation on the analysis, I applied variation partitioning by RDA (Borcard *et al.* 1992) to the matrix, using fish distribution as the dependent matrix, and environmental and spatial predictors as two sets of predictors. Variation partitioning is used to identify common and unique contributions to model prediction and hence better address the question of the relative influences of the groups of independent variables considered in the model. When partitioning variation in RDA, independent variables are grouped into sets

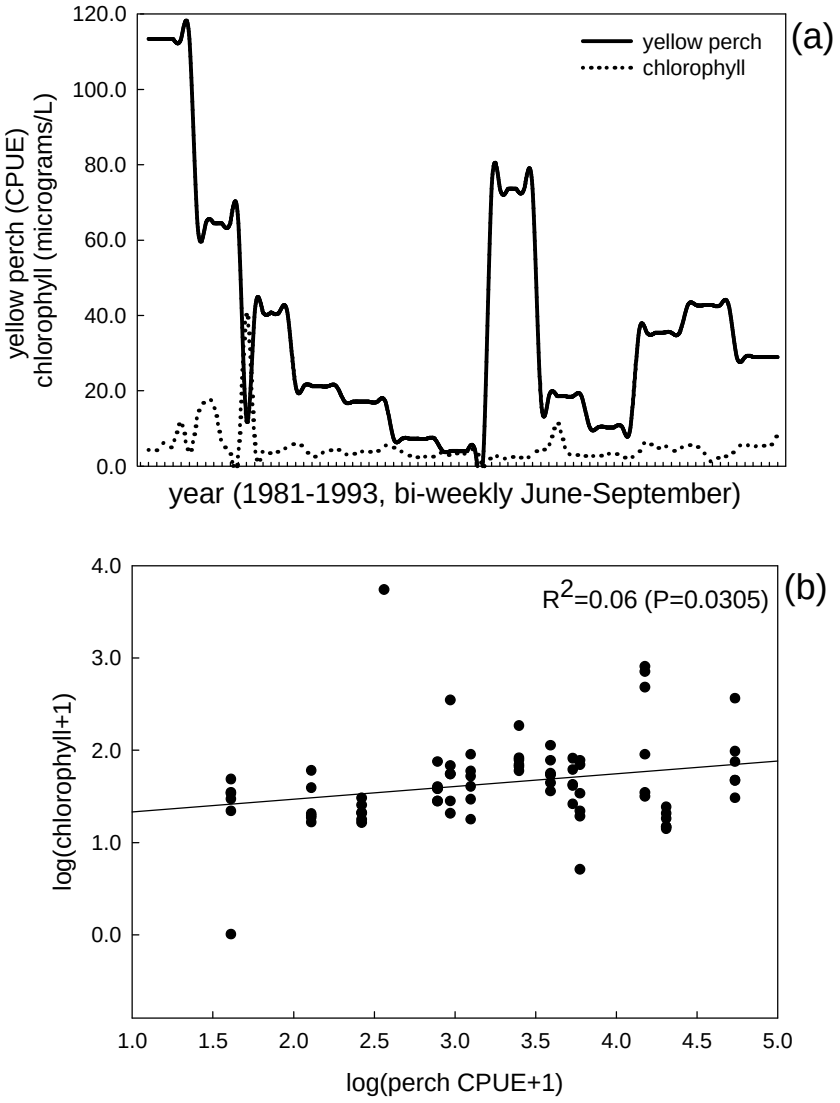



Figure 5. (a) Time series of yellow perch catches and chlorophyll concentration; (b) linear regression of chlorophyll on yellow perch based on the time series values.

representing broad factors (here environment and space). In that context, variation partitioning is more suitable than analyzing the individual contributions of regressors via their partial correlation coefficients. In this approach, the total adjusted percentage of variation explained by the model is partitioned into unique and common contributions of the sets of predictors (details of these calculations are provided in Peres-Neto *et al.* 2006).

A forward stepwise selection procedure for RDA (ter Braak & Smilauer 2002) was applied to select spatial eigenvectors that are important in explaining the spatial autocorrelation in species distribution. Only four out of the original 12 eigenvectors representing positive spatial autocorrelation were retained. Results of the variation partitioning based on the selected eigenvectors and adjusted fractions of variation are presented in Fig. 6. After partitioning the spatial variation, environment continued to explain a significant amount of the variation (fraction [a] in Fig. 6).

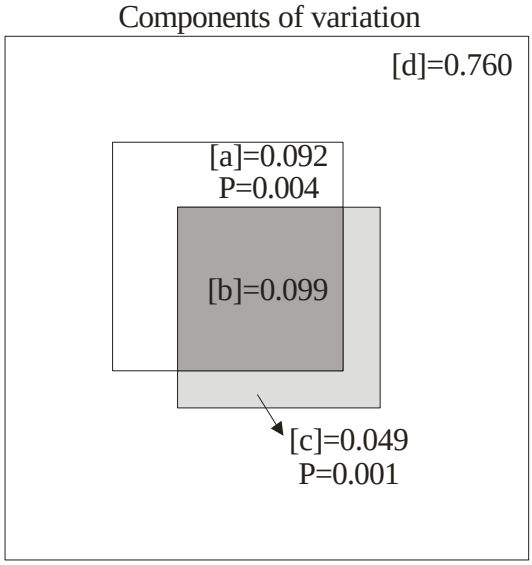



Figure 6. Variation partitioning Venn diagrams representing the adjusted percentages of unique contribution of spatial [c] and environmental [a] components to the fish distribution in Macacu river.

Significance of fractions was tested using permutation tests (999 permutations were applied here; see Legendre & Legendre 1998 for details on test of fractions). As applied here, the variation partitioning fractions presented in Fig. 6 represent the following values: fraction [a] is the unique variation of species distribution explained by environment (i.e., after the spatial variation is removed from the environmental variation); fraction [c] is the unique variation of species distribution explained by spatial autocorrelation (i.e., after the environmental variation is removed from the spatial predictors); fraction [b] is the variation in species distribution that cannot be uniquely assigned to space or environment and represents the level of multicolinearity (i.e., shared variation) between the two sets of predictors (i.e, space and environment). Only fractions [a] and [c] can be tested.

Variation partitioning raises an important issue related to the control of spatial autocorrelation in any ecological model, and not only in the present case. It is not all variation due to spatial autocorrelation that will bias statistical inference, but only a particular component, here expressed by fraction [b]. Unlike in variation partitioning, fraction [b] cannot be easily estimated for some other modeling procedures (e.g., logistic regression). However, this fraction is automatically considered in significance tests via partial approaches that eliminate multicolinearity among predictors when spatial eigenvectors are entered in the model, in addition to the other regressors of interest. This point is not commonly understood and is worthy of mention because not all spatial autocorrelation is detrimental to interpretation. In fact, the left-over spatial component (i.e., residual variation of spatial variation independent of environment, or fraction [c] in ordinary least square models) can be interpreted further and aid in understanding contagious ecological processes that are important in driving species distribution (e.g., dispersal). However, it is important to keep in mind that part of this residual variation may still represent spatial autocorrelation of unmeasured environmental predictors.

## *ANALYSIS OF A PHYLOGENETIC DATA*

Here an ecomorphological data set comprising 13 species of West Indian *Anolis* lizards (Losos 1990; data is presented in his Table 1) is analyzed to test whether morphological differences amongst species result in differences in performance in this group of lizards. This hypothesis was tested by Losos (1990) using Felsenstein's contrasts (1985) for controlling the effects of phylogeny and canonical correlation analysis. For the sake of illustration, I test the same hypothesis

but using the eigenvector method and variation partitioning for RDA instead.

The phylogeny and phylogenetic covariance matrix are presented in Fig. 3. There was no evidence that morphology and performance (values were ln-transformed) are phylogenetically dependent (morphology: = -0.011, P = 0.3260; performance = -0.002, P = 0.3720), suggesting that the association between the two multivariate traits (= 0.9017, P = 0.001) is due to adaptive evolution where convergence evolution has been widespread in these lizards (see Losos 1990 for further discussion).

## CONCLUDING REMARKS

My goals here were two fold: (1) expose ecologists to the problem related to statistical inference when testing the association between data sets that are autocorrelated and (2) introduce a relatively new class of predictors based on eigenfunction analysis that can be easily incorporated in analytical approaches for controlling the bias due to autocorrelation. In addition, I have shown the flexibility of this class of methods to the types of data in which ecologists are currently most interested, namely temporal, spatial and phylogenetic analyses. Here I applied two commonly used methods based on ordinary least-square regression (RDA and regression) as means of testing the association between variables and data sets. However, the two method introduced here for controlling autocorrelation is readily extensible to any type of modelling procedure such as GLMs (e.g., logistic/binomial and Poisson regressions) and modern computational procedures (see Elith *et al.* 2006 for a review) by using eigenvectors as additional predictors in these models. It is worth of noting that the autoregressive approach used in many temporal, spatial and phylogenetic applications becomes cumbersome and difficult to estimate when extended to GLMs and to other modelling techniques. The principal problem with autoregression is the normalizing constant, which is the Jacobian term in the linear model case, and requires Markov chain Monte Carlo estimation in the GLM case. In addition, there is nothing akin to the autoregressive model for modelling multiple variables (e.g., multiple species, several morphological characters) as in canonical analysis (e.g., RDA).

I feel that is particularly important to stress the point that not all the autocorrelation present in data that will affect statistical inference and is important to control only the component that bias inference. Thus, future analytical developments should attempt to separate the autocorrelation bias component from the one that may prove of interest for understanding important ecological processes such as contagious processes driving spatial patterns in species distributions. Parts of variation in data due to these contagious processes may cause bias and should be controlled for, but the component that does not bias statistical interpretation should be kept and further analyzed. This is relatively simple to perform by conducting residual analysis where the common component of variation due to autocorrelation in the data sets being analyzed is removed and the independent component is analyzed. Another important point that should be reiterated is that analytical approaches for controlling the effects of autocorrelation should be only considered in the presence of dependence among observations and that is the reason we should start by testing whether or not the data sets involved are autocorrelated. Abouheif (1998) raised the point in evolutionary ecology that if analytical approaches are used to control for phylogenetic autocorrelation when it is not relevant, that may also introduce bias to the analysis. Another important point is that statistical inference is only affected when both variables (or data sets) are autocorrelated. If only one variable is autocorrelated whereas the other is not, then the analysis is not affected. This also reiterates the point of removing only the common component of variation stated above.

Model selection can play an important role in controlling the effects of autocorrelation. For each eigenvector that is considered in the model, the model is penalized by the loss of degrees of freedom when testing for the association between the original variables of interest (e.g., species distributions and environmental characteristics). Therefore, only a subset of the relevant eigenvectors should be considered, chosen by a model selection procedure, in order to maximize the chances of detecting true associations between the variables involved (i.e., increase power of the test). Modelling selection procedures may bias the estimation of variable contribution and model fit, among others, but if not used, there can be a reduction in the power of statistical procedures to test the association

between data sets. Therefore, this problem opens another avenue for future research. Griffith & Peres-Neto (1996) advocated for a modelling selection procedure that minimizes the autocorrelation in residuals of GLM models, however the advantages and disadvantages of this and other model selection procedures should be further examined. In this contribution, I only considered eigenvectors that are positively autocorrelated, but future research should investigate whether negative autocorrelation also promote bias in statistical inference. If that is the case, then the method can also accommodate this type of autocorrelation by using the eigenvectors that represent negative autocorrelation in the analysis (i.e., eigenvectors with negative eigenvalues or MI). There are many possible avenues for applying and expanding the applications of eigenvector predictors, and the present study is an illustration of how this relatively new and flexible technique can be used in ecological analysis.

Acknowledgments - Funding from NSERC Discovery Grants is gratefully acknowledged. I would like to thank Daniel Borcard, José-Alexandre F. Diniz-Filho, Stéphane Dray, Daniel Griffith, Donald Jackson and Pierre Legendre for the countless discussions on the issues of autocorrelation in ecology and evolution.

## REFERENCES

ABOUHEIF, E. 1999. A method for testing the assumption of phylogenetic independence in comparative data. *Evolutionary Ecology Research*, 1: 895-909.

BAVAUD, F. 1998. Models for spatial weights: A systematic look. *Geographical Analysis*, 30: 153-171.

BEISNER, B.E.; IVES, A.R. & CARPENTER, S.R. 2003. The effects of an exotic fish invasion on the prey communities of two lakes. *Journal of Animal Ecology*, 72: 331-342.

BENCE, J.R. 1995. Analysis of short time series: correcting for autocorrelation. *Ecology*, 76: 628-639.

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.

BUTLER, M.A.; SCHOENER, T.W. & LOSOS, J.B. 2000. The relationship between habitat type and sexual size dimorphism in Greater Antillean *Anolis* lizards. *Evolution*, 54: 259-272.

CLARK, J.S. & BJØRNSTAD, O. 2004. Population time series: Process variability, observation errors, missing values, lags, and hidden states. *Ecology*, 85: 3140-3150.

CLIFFORD, P.; RICHARDSON, S. & HÉMON, D. 1988. Assessing the significance of the correlation between two spatial processes. *Biometrics*, 45: 123-134.

DALE, M.R.T. & FORTIN, M.J. 2002. Spatial autocorrelation and statistical tests in ecology. *Écoscience*, 9: 162-167.

DENNIS, B. & TAPER, M.L. 1994. Density dependence in time series observations of natural populations: estimation and testing. *Ecological Monographs*, 64: 205-224.

DINIZ-FILHO, J.A.F. 2000. *Métodos Filogenéticos Comparativos*. Holos, Riberão Preto.

DINIZ-FILHO, J. A. F. 2001. phylogenetic autocorrelation under distinct evolutionary processes. *Evolution*, 55: 1104-1109.

DINIZ-FILHO, J.A.F. & BINI, L.M. 1996. Assessing the relationship between multivariate community structure and environmental variables. *Marine Ecology Progress Series*, 143: 303-306.

DINIZ-FILHO, J.A.F.; SANT'ANA, C.E.R. & BINI, L.M. 1998. An eigenvector method for estimating phylogenetic inertia. *Evolution*, 52: 1247-1262.

DOUGLAS, M.E. & ENDLER, J.A. 1982. Quantitative matrix comparisons in ecological and evolutionary investigations. *Journal of Theoretical Biology*, 99: 777-795.

DOUGLAS, M.E. & MATTHEWS, W.J. 1992. Does morphology predict ecology? Hypothesis testing within a freshwater stream fish assemblage. *Oikos*, 65: 213-224.

DOUGLAS, M.E.; MINCKLEY, W.L. & DEMARAIS,B.D. 1999. Did vicariance mold phenotypes of western North American fishes? Evidence from *Gila* river Cyprinids. *Evolution*, 53: 238-246.

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.

ELITH, J.; GRAHAM, C.H.; ANDERSON, R.P.; DUDI´K, M.; FERRIER, S.; GUISAN, A.; HIJMANS, R.J.; HUETTMANN, F.; LEATHWICK, J.R.; LEHMANN, A.; LI, J.; LOHMANN, L.G.; LOISELLE, B.A.; MANION, G.; MORITZ, C.; NAKAMURA, M.; NAKAZAWA, Y.; OVERTON, J.M.; PETERSON, A.T.; PHILLIPS, S.J.; RICHARDSON, K.; SCACHETTI-PEREIRA, R.;

SCHAPIRE, R.E.; SOBERO´N, J.; WILLIAMS, S.; WISZ, M.S. & ZIMMERMANN, N.E. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29: 129-151.

FELSENSTEIN, J. 1985. Phylogenies and the comparative method. *American Naturalist*, 125: 1-15.

FELSENSTEIN, J. 1988. Phylogeny and quantitative characters. *Annual Review of Ecology and Systematics*, 19: 445-471.

FRECKLETON, R. P.; HARVEY, P. H. & PAGEL, M. 2002. Phylogenetic Analysis and Comparative Data: A Test and Review of Evidence. *American Naturalist*, 160: 712–726.

GARLAND JR., T. & A. R. IVES. 2000. Using the past to predict the present: Confidence intervals for regression equations in phylogenetic comparative methods. *American Naturalist*, 155: 346-364.

GARLAND JR., T.; HARVEY, P. H. & IVES, A.R. 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. *Systematic Biology*, 41: 18-32.

GRIFFITH, D. 2000. A Linear Regression Solution to the Spatial Autocorrelation Problem. *Journal of Geographical Systems*, 2: 141-56.

GRIFFITH, D.A. & PERES-NETO, P.R.. 2006. Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses in exploiting relative location information. *Ecology* ( *in press*).

HARVEY, P.H.; MARTIN, R.D. & CLUTTON-BROCK, T.H. 1987. Life histories in comparative perspective. Pp 181-196. *In*: B.B. Smuts, D.L. Cheney, R.M. Seyfarth, R.W. Wrangham & T.T. Struhsaker, (eds.), Primate Societies. University of Chicago Press, Chicago.

IVES A.R. & ZHU, J. 2006. Statistics for correlated data: phylogenies, space, and time. *Ecological Applications*, 16: 20-32.

JACKSON, D.A. & HARVEY, H.H. 1993. Fish and benthic invertebrates: community concordance and community-environment relationships. *Canadian Journal of Fisheries and Aquatic Sciences*, 50: 2641-2651.

JENKINS, D.G. & BUIKEMA, A.L. 1998. Do similar communities develop in similar sites? A test with zooplankton structure and function. *Ecological Monographs*, 68: 421-443.

LEGENDRE, P. 1993. Spatial autocorrelation: Trouble or new paradigm? *Ecology*, 74: 1659-1673.

LEGENDRE, P. & FORTIN, M.J. 1989. Spatial pattern and ecological analysis. *Vegetatio*, 80: 107-138.

LEGENDRE, P. & LEGENDRE, L. 1998. *Numerical Ecology*. (English Second Edition Elsevier Science BV, Amsterdam.

LEGENDRE, P. & ANDERSON, M.J. 1999. Distance-based redundancy analysis: testing multi-species responses in multi-factorial ecological experiments. *Ecological Monographs*, 69: 1-24.

LEGENDRE P.; BORCARD, D. & P.R. PERES-NETO. 2005. Analyzing beta diversity: partitioning the spatial variation of community composition data. *Ecological Monographs*, 75: 435-450.

LEGENDRE, P.; DALE, M. R. T.; FORTIN, M.J.; CASGRAIN, P. & GUREVITCH, J. 2004. Effects of spatial structures on the results of field experiments. *Ecology*, 85: 3202-3214.

LOSOS, J.B. 1990. Ecomorphology, performance capability, and scaling of West Indian *Anolis* lizards: an evolutionary analysis. *Ecological Monographs*, 60: 368-388.

MANLY, B.J.F. 1997. Randomization, Bootstrap and Monte Carlo Methods in Biology (Second Edition). Chapman and Hall, London.

MARTINS, E. P.; DINIZ-FILHO, J.A. & HOUSWORTH, E.A. 2002. Adaptation and the comparative method: A computer simulation study. *Evolution*, 56: 1-13.

MARTINS, E. P. & GARLAND JR., T.1991. Phylogenetic analyses of the correlated evolution of continuous characters: a simulation study. *Evolution*, 45: 534-557.

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. 2004. Patterns in the co-occurrence of stream fish metacommunties: the role of site suitability, morphology and phylogeny versus species interactions. *Oecologia*, 140: 352-360.

PERES-NETO, P.R. & OLDEN, J.D. 2001. Assessing the robustness of randomization tests: examples from behavioural studies. *Animal Behaviour*, 61: 79-86.

PERES-NETO, P.R.; OLDEN, J.D. & JACKSON, D.A. 2001. Environmentally constrained null models: site suitability as occupancy criterion. *Oikos*, 93: 110-120.

PERES-NETO, P.R.; LEGENDRE, P.; DRAY, S. & BORCARD, D. 2006. Variation partitioning of species data matrices: estimation and comparison of fractions. *Ecology* (*in press*).

RAO, C.R. 1964. The use and interpretation of principal component analysis in applied research. *Sankhyaá*, Ser. A, 26: 329-358.