Journal of Biogeography (J. Biogeogr.) (2012) 39, 1–9

GUEST EDITORIAL

Eight (and a half) deadly sins of spatial analysis

Bradford A. Hawkins

Department of Ecology & Evolutionary Biology, University of California, Irvine, CA 92697, USA

Correspondence: Bradford A. Hawkins, Department of Ecology & Evolutionary Biology, University of California, Irvine, CA 92697, USA.
E-mail: bhawkins@uci.edu

ABSTRACT

Biogeography is spatial by nature. Over the past 20 years, the literature related to the analysis of spatially structured data has exploded, much of it focused on a perceived problem of spatial autocorrelation and ways to deal with it. However, there are a number of other issues that permeate the biogeographical and macroecological literature that have become entangled in the spatial autocorrelation web. In this piece I discuss some of the assumptions that are often made in the analysis of spatially structured data that can lead to misunderstandings about the nature of spatial data, the methods used to analyse them, and how results can be interpreted.

Keywords

Biogeographical analysis, diversity gradients, geographical ecology, macroecology, multiple regression, regression trees, spatial analysis, spatial autocorrelation, spatial regression.

INTRODUCTION

First things first. Some readers will recognize that the title of this piece is similar to a recent article by Robert P. Freckleton on seven common problems encountered in comparative analysis (Freckleton, 2009). I borrow his title for reasons beyond a limited imagination. Although phylogenetic and spatial data are not identical in causes and structure, the analysis of data containing autocorrelation in spatial and phylogenetic contexts shares many similarities (Peres-Neto, 2006), and many of the concepts and methods in comparative biology are applicable to spatial ecology and biogeography, and vice versa [see, for example, Gittleman & Kot's (1990) extension of Moran's *I* and spatial correlograms to a phylogenetic framework]. The issues discussed by Freckleton (2009) also apply to the topic of this essay, although I attempt to minimize overlap as much as possible. Even so, workers conducting analyses of spatially structured data should read his essay. The points of view differ, but his discussion of broad issues frees me to focus on more specific issues related to spatially structured data.

For many geographical ecologists and biogeographers the spectre of spatial autocorrelation was released by Legendre (1993). Prior to that, studies of diversity gradients (e.g. Rabinovich & Rapoport, 1975; Schall & Pianka, 1978; Pianka & Schall, 1981; Currie & Paquin, 1987) were unconcerned with the fact that the same species appeared more than once in the data set and that this generates a lack of independence in plot or grid-cell richness estimates. Because species are not randomly distributed in the environment (for a number of reasons), the occurrences of species in multiple cells or plots generates the autocorrelation in geographical data not due to data quality problems. And lest we believe that autocorrelation, spatial or otherwise, is restricted to geographical ecology and biogeography, it should be remembered that most experimental ecology conducted in the field or laboratory in which an assemblage-level metric such as species richness is either a predictor or a response variable will contain multiple species occurrences. We can imagine the difficulty that D. Tilman would have had in his now classic experiments linking species and functional richness to ecosystem function (Tilman & Downing, 1994; Tilman *et al.*, 1996, 1997) if interpretation depended on counting each plant species in only one of his plots. Indeed, under this constraint it is difficult to see how an ecologist could study any aspect of diversity either experimentally or observationally, and the concept of beta diversity quickly loses its meaning. Given the ubiquity of autocorrrelated data in ecology and biogeography, does it follow that the entire literature is flawed? If not, what is all the fuss about, and is a fixation on spatial autocorrelation overshadowing more serious issues related to the evaluation of geographical patterns?

What follows is a discussion of nine claims often made in papers analysing geographical patterns, or assumptions that

Journal of Biogeography

may influence the choice of analytical methods or interpretation of results. I also include a few examples of how they might be dealt with. I do not cite many papers containing the problems on which I focus, as I am not trying to pick a fight with individual researchers. I am also not in a position to cast stones. Finally, my evaluations are not the only possible point of view. There is room for disagreement about many aspects of spatial analysis, and the structure of spatial data is complex. On the other hand, these issues deserve attention, and although almost everything I say has been said more than once in the statistical or ecological literature, all of these problems continue to plague analyses of geographical gradients of taxon richness, body size, range size, and all other assemblage-level metrics generated by biogeographers and macroecologists. We may not all agree on the solutions, but the field will benefit if we all at least recognize the problems.

SPATIAL AUTOCORRELATION GENERATES BIAS

Nature is autocorrelated [Tobler's (1970, p. 236) first law of geography, 'everything is related to everything else, but near things are more related than distant things']. In our context, all species are distributed non-randomly across a wide range of spatial scales, even if the distributions of small groups of individuals may sometimes be indistinguishable from random at scales of a few metres to hectares. As argued by Legendre (1993), the existence of spatial autocorrelation is not *bias* or *artefact*, it is what biogeographers and geographical ecologists want to understand. Given the spatially structured distributions of species arising from biological processes operating in a spatially patterned environment, any set of samples or representations of nature must also contain this structure if it is accurate. What this clearly means is that broad-scale data that do not contain spatial structure (which is what most workers mean when they say spatial autocorrelation) are not an accurate description of nature, or they are missing key information that limits their value for understanding the spatial patterns we are studying. If spatial autocorrelation is part of nature, and we are trying to understand nature, it makes little sense to claim that spatial autocorrelation in data represents some sort of bias, artefact or distortion. The confusion arises because of an epistemological framework that confounds Type I error rates associated with the use of frequentist statistics, attempts to use multiple regression for causal inference, data quality issues (a separate problem I will come to later), and the belief that we have to model everything to understand anything. There has also been some arguing at cross-purposes by a failure to always articulate clearly the difference between autocorrelation as a pattern in nature versus autocorrelation in residuals of statistical models. There are a large number of overlapping problems here, which is why it has proven difficult to sort out the issues in the literature. I discuss these in subsequent sections but begin by focusing the argument to make the issues tractable in a relatively short space.

The most common method for analysing geographical gradients is linear regression, whether ordinary least squares (OLS) or spatially explicit. Related methods are sometimes used, but most linear modelling approaches hope to *test* hypotheses for diversity gradients, generate unstandardized and/or standardized regression coefficients, rank standardized coefficients by their *importance* when the analysis includes multiple predictors, and *explain* variance/deviance. For now, I will restrict the discussion to parametric multiple regression, as the structural limitations in this method are major causes of the disagreement about the analysis of spatial data.

One aspect of spatial autocorrelation that is often repeated and universally accepted is that it may result in a lack of independence among residuals, leading to an underestimation of standard errors and an overestimation of significance arising from the standard significance tests used in parametric regression. This is not controversial, and there is an obvious solution: stop doing significance tests. There are also a number of more complicated solutions to this problem, but they do not deal with the fundamental issue. Classical null hypothesis testing is not the appropriate way to deal with the types of data we analyse (Quinn & Dunham, 1983; Burnham & Anderson, 2002), and many workers are adopting an alternative analytical framework that does away with generating *P*-values from a comparison of observed data against null models, although even this is not a perfect solution from a statistical standpoint. I will treat this in a subsequent section. Significance testing and null models aside, a much more serious issue is the claim that regression coefficients are biased by spatial autocorrelation, which can only be remedied by use of spatially explicit methods. This is not true in one sense and is arguable in another.

The first thing to note is that OLS regression coefficients, except with respect to estimating standard errors of coefficients, are not biased by spatial autocorrelation in a statistical sense (Cressie, 1993; Dutilleul, 1993; Fortin & Dale, 2005; Schabenberg & Gotway, 2005; Diniz-Filho *et al.*, 2007; Hawkins *et al.*, 2007). Some workers are concerned with precision in OLS, which if they run simulations can cause some sample slopes to be far from the true slope by chance. However, this will only occur with small sample sizes and is actually a problem of insufficient data due to poor study design rather than a problem arising from spatial autocorrelation, which is why throwing out data because of autocorrelation is not recommended. But even though the precision claim is true, it is not bias in any sense of the word, although workers who claim that OLS is biased are often confounding bias and precision. The point is, statisticians widely agree that OLS regression slopes are unbiased estimates in the face of spatial autocorrelation. Given this, why do claims that one needs to control for spatial autocorrelation to generate unbiased interpretations of regression models persist in the ecological literature? Again, I believe it arises from a confounding of multiple issues. If all regression methods generated identical models for any particular data set there would be no arguments. But they sometimes do not, and then the most

common claim is that it must represent the bias of spatial autocorrelation. But if OLS regression really is statistically unbiased, differences among methods must have alternative, and perhaps multiple, origins. The following sections discuss some of those sources of uncertainty in analyses of geographical gradients, most of which focus on the real problem, multiple regression, whether spatial or not.

SPATIAL REGRESSION IS BEST

This assumption arises directly from the supposition that OLS regression must be biased. But the claim that spatial regression is the solution bears closer scrutiny. Bini *et al.* (2009) recently compared OLS and eight spatially explicit methods for 97 broad-scale ecological data sets and found that when faced with real data multiple regression is a mess, whether spatially explicit or not. The problem is that models may (or may not, in most cases) contain substantially different regression coefficients for any particular data set across methods, but predicting when methods will generate different models or identifying why results differ in some data more than in others is extremely difficult. When different methods give different results, and we do not understand why, blanket claims concerning the merits of one method over another are untenable. As far as I know, all claims that regression methods that model space as a process [generalized least squares (GLS) and autoregression] are better than OLS with respect to parameter estimation derive from simulated data generated with unrealistic assumptions and contrived structures, none of which accurately capture the structure of real geographical data. Although these simulation studies are useful for evaluating hypothetical situations from a statistical standpoint, it is not valid to extend results from a simulation into situations where data have completely different spatial structures, covariance patterns and spatial dependencies, as these characteristics can violate underlying assumptions of ALL multiple regression methods. After all, it is not difficult to generate simulated data to demonstrate that parametric methods are more powerful than nonparametric methods, but that does not justify the use of parametric statistics on all data. If it did, nonparametric statistics would not exist, and in general there would be far fewer statistical methods. As a note, Bini *et al.* (2009) focused on spatially explicit regression methods, but the same issues apply as well to spatially implicit methods that use weights rather than explicit spatial predictors (P. R. Peres-Neto, Université du Quebec à Montréal, pers. comm.).

There are a number of ways to come to grips with the inherent uncertainly associated with all forms of multiple regression, which is the primary source of what some refer to as bias. First, admit it. Presenting a single model as if it is the only solution is not the best strategy. A slightly better approach is to compare OLS and a spatial model to quantify the similarity among the results provided by these models, but given the unpredictable nature of how ecological data respond to different regression methods, deciding which methods to compare is itself a problem (Bini *et al.*, 2009). At the other extreme, reporting nine or 10 sets of models using all possible methods is not necessarily helpful, as it greatly complicates analyses and probably generates more confusion than it resolves. One can use model averaging, which does have real advantages (Burnham & Anderson, 2002; Diniz-Filho *et al.*, 2008), but this does not resolve the issue of whether the particular methods being used are even valid, which brings us to the problem of uncritical uses of multiple regression.

THE WORLD IS STATIONARY

All parametric linear regression methods assume that relationships among predictor and response variables are invariant throughout the entire span of the data. This is the assumption of stationarity. In the case of OLS, violation of this assumption is not critical as long as it is remembered that the model coefficients represent an *average* of relationships across the entire data set and will not apply to subsets of data. But spatially explicit regression methods are also sensitive to this assumption, and this can generate serious problems. Spatial autoregression shifts the interpretation of the coefficients from global to more local scales (Fotheringham *et al.*, 2002; Diniz-Filho *et al.*, 2007), and most spatial regression methods generate a single, locally adjusted *global* estimate of each coefficient, which in the face of non-stationarity can be misleading. That is, in non-stationary data the single, *semi-local* regression coefficients (*sensu* Fotheringham *et al.*, 2002) generated by spatial regression cannot be interpreted with confidence because a unique local coefficient does not exist in the data. Much like spatial autocorrelation itself, this fundamental property of geographical data was largely ignored until recently (Foody, 2004; Bickford & Laffan, 2006; Cassemiro *et al.*, 2007; Beale *et al.*, 2010; Landeiro & Magnusson, 2011), and it remains very uncommon for geographical ecology papers to report that they have determined whether their data are stationary or not. This is despite the fact that non-stationarity is common in broad-scale data; e.g. of the 97 data sets compiled by Bini *et al.* (2009), in 45 the difference in Akaike's information criterion (AIC) values between a geographically weighted regression and all other spatial and non-spatial model types was > 2, in most cases much more, indicating non-stationarity. If these numbers are representative it indicates that the results of up to half of the published analyses using spatial regression on large-extent ecological data require confirmation based on this issue alone.

The minimal solution to this problem is that whenever an analysis uses parametric linear regression a clear statement is provided that the stationarity of the data has been tested. There are a number of methods to detect non-stationarity in space, including comparing means and variances of variables in sliding windows of different size (Fortin & Dale, 2005), pocket plots (Cressie, 1993), or a comparison of the residual sums of squares of OLS versus geographically weighted regression models (Fotheringham *et al.*, 2002). If one of these tests identifies non-stationarity and authors want to use an autoregressive method [autoregressive moving average

(ARMA), conditional autoregression (CAR), simultaneous autoregression (SAR), etc.], they must inform readers that a basic assumption of the analysis has been violated. Users of OLS regression should also warn readers that the global coefficients from their model(s) may be hiding local variation in relationships. It is possible to model non-stationarity using high-order autoregression (Galbraith & Zinde-Walsh, 1997), but this statistical fix does not resolve the underlying problem of ecological inference. A much better alternative is to use non-stationarity to investigate potential underlying reasons for it (e.g. Hortal *et al.*, 2011). In this sense non-stationarity is like spatial autocorrelation; it is not a problem to be eliminated or modelled away statistically but is a valuable source of information for understanding nature.

Finally, one can include an analytical method that can capture non-stationary relationships, such as a regression tree or its derivatives. Classification and regression tree (CART) analysis (Breiman *et al.*, 1998) is a machine-learning method that is ideally suited to geographical analysis. It works using any combination of continuous and categorical variables, it makes no assumptions about the shapes of relationships, and it can detect changes in relationships across space. Similar to multiple regression, variables can be ranked using variance improvement values to estimate the explanatory power of each variable across the entire tree, and variable importance values identify levels of *masking* of one variable by others, facilitating interpretation in the presence of collinearity (a type of höchst model averaging). CART is making inroads in all branches of ecology due to its versatility and power (a Web of Knowledge topic search in September 2011 using 'regression tree' AND 'ecology' generated 63 hits; 47 published after 2005), and it is beginning to be used in biogeography and geographical ecology (e.g. Wilson, 2008; Bachraty *et al.*, 2009; Fink *et al.*, 2010; Hawkins, 2010).

PARTIAL REGRESSION COEFFICIENTS MEAN SOMETHING

This is perhaps the major reason why workers worry about which regression method to use. The ability to say that we have identified the most *important* influence on a spatial ecological gradient provides a strong stimulus to rank standardized partial regression coefficients and assume that they are informative. But it is unlikely that they have real explanatory power in the complicated types of data we analyse, as discussed by Grace & Bollen (2005, p. 287):

> While investigators commonly ask, 'What is the relative importance of a set of causes controlling some observed phenomenon?' we must conclude that when predictor variables are correlated for unknown reasons, standardized partial regression coefficients do not provide an answer to this question…Multiple regression, which is inherently designed to ignore the causes behind the correlations among a set of predictors, makes for a particularly poor approach to understanding.

The many problems with multiple regression when applied to observational data sets with complex covariance structures, unmeasured collinear driving variables, and a lack of experimental controls are well known but continue to be widely ignored, largely because of historical inertia, the fact that the method is known by virtually everyone, and it is easy to do. While attempts have been made to remind workers of some of the pitfalls associated with multiple regression in an ecological context and to propose methods to deal with them (Freckleton, 2002, 2009; Graham, 2003; Whittingham *et al.*, 2006), it remains that while it is a reasonable method for predicting relationships, it is a very poor one for explaining them. And if no partial regression coefficient has explanatory power in an epistemological sense, then arguments about the value of a coefficient derived from this or that form of regression rapidly fall into the realm of arguing about numbers of angels dancing on the heads of pins.

Although multiple regression continues to represent the most common form of analysis, alternatives are being increasingly applied. For example, as mentioned above, classification and regression trees are now being used, and they have many clear advantages over multiple regression (Breiman *et al.*, 1998). Workers who find it difficult to admit that correlations are not processes (see next section) may think CART is too descriptive, but it can be used in conjunction with parametric linear modelling for comparison and improved interpretation of results. Structural equation modelling (SEM) is also a powerful tool (Shipley, 2000), although interpretation also requires care (Grace & Bollen, 2005). In many situations it is difficult to generate cause-and-effect hypotheses about relationships among multiple environmental predictors, but exploratory methods for SEM have been developed that allow one to identify potential models consistent with the data, relaxing to some extent the need to have an *a priori* set of hypotheses for causal links in complex data sets (Shipley, 2000). In addition to these two suggestions, there are a number of statistical fixes that can be used to deal with specific problems with multiple regression, but space is limited. The point of this section is simply to reinforce many previous reminders in the literature that multiple regression is not a good method for understanding relationships between diversity (or any other assemblage-level metric) and the environment, and quibbles over which particular brand of regression is better or worse is a distraction that cannot resolve the basic problems we face in the field.

REGRESSION COEFFICIENTS IDENTIFY EFFECTS

The phrase 'correlation does not imply causation' must be the most widely known statistical aphorism in ecology, and the most often ignored (with apologies to Shipley, 2000). This is another reason why workers argue about which form of regression is appropriate; if regression coefficients are linked to particular processes, then knowing which of several candidates is most strongly correlated with a geographical gradient helps us evaluate processes. But surely everyone understands that a correlation, however strong, is not evidence that a particular process is operating. Perhaps the closest we can come

to evaluating a process is when it makes a unique prediction about a relationship between some variable and diversity, but the predicted relationship is not found. However, even then it can be argued that the failure to fit the prediction is because of the effect of some unmeasured process obscuring the expected pattern for the variable we tested. It is the ubiquity of potential collinear variables and processes and the impossibility of measuring every relevant variable at every scale that undermines interpretation of standardized regression coefficients under any circumstance. A species richness gradient along a north–south transect may generate a larger partial regression coefficient for latitude than for any environmental variable, but knowing this does not tell us anything about any particular effect. On the other hand, it cannot be denied that the allure of referring to regression coefficients as effects is so strong that even seasoned statisticians sometimes do it (Ord, 1975). It is one thing to hypothesize *a priori* that a correlation reflects a particular process, but it is another to use language that assumes it.

The term *effect* is used and accepted with respect to SEM and meta-analysis (which measures *effect sizes*), but in the former the philosophy underlying the method explicitly treats these effects as hypotheses, and in the latter the term is used in a statistical sense rather than a mechanistic one. Admittedly, the use of inappropriate language with respect to regression is not a *deadly sin* (*sensu* Freckleton, 2009), and hence the half a sin in the title of this essay, but the loose use of the word *effect* when based solely on a correlation/regression coefficient can reflect woolly thinking and suggest to colleagues the lack of an understanding of scientific inference.

SPECIES RICHNESS GENERATES BIAS

Geographical analyses of assemblage-level attributes other than taxon richness are sometimes claimed to be influenced by underlying richness patterns. The typical argument is that averages of, e.g. range size, body size or some phylogenetic metric, are potentially biased by variation in the numbers of species in different places, so that the means have to be corrected for sample size differences. However, this represents a fundamental misunderstanding of sampling theory. Given a distribution of any variable, all random samples of any size are unbiased estimators of the parametric mean of that variable. If this were not true, sample-based statistics would be impossible. As explained in introductory textbooks on statistics, any estimate of a mean will converge on the parametric mean irrespective of the number of units (species) sampled *if the sample is random*. Of course, if the species in a place are a non-random sample of all possible species then averages will indeed vary spatially, but this is the biology we are trying to understand and does not require correction. As is the case with complaints about OLS regression, the likely source of this claim is a confusion between precision and bias. But an understanding of the nature of this problem does not require a sophisticated knowledge of statistics, and it is clear that the claim that richness generates bias in estimates of means is without foundation.

THE EARTH IS ROUND ($P < 0.05$)

This section heading is a pilfered title (Cohen, 1994) that expresses the issue succinctly. Even a cursory reading of the literature indicates that the field is generally following Burnham & Anderson's (2002) advice and moving from frequentist statistics to an information-theoretic approach that focus on model selection rather than on tests of null models. The advantages of this are well known and hopefully do not need belabouring. In our field it also represents a positive step towards the solution to the problem of inflated Type I error rates when residuals are not independent. This does not mean that AIC values can be taken at face value when the models do not capture all spatial pattern in the data (Diniz-Filho *et al.*, 2008). Nor can it resolve the more fundamental problem that models may not include all of the variables needed for ecological and evolutionary interpretation: but that is a research problem rather than an analytical one.

The concerns of Diniz-Filho *et al.* (2008) that AIC does not necessarily identify the minimally adequate model are valid, but in practice is a regression model with six variables accounting for 76.2% of the variance in taxon richness really better or worse than a four-variable model accounting for 75.9%, whatever the AIC value? Most workers would probably agree that one of the major advantages of model selection methods is that multiple alternatives are evaluated, and readers can judge for themselves how the various options compare in terms of explanatory power and differences in relative complexity. Including weights for variables and models and model averaging also aids evaluation. For those worried about the tendency of AIC to select overparameterized models, the use of the Bayesian information criterion (BIC) provides an alternative for multimodel inference (Link & Barker, 2006).

A small issue with the switch from null-hypothesis based methods to model selection is that they are philosophical alternatives (Burnham & Anderson, 2002). That is, one should either use *P*-values or AIC/BIC values to evaluate models. Because virtually all statistical packages now output both types of statistics, some workers interpret this to mean that they are complementary and use both for model evaluation. This should be avoided if the reason for using model selection methods is based on the realization that significance tests are not appropriate in most geographical contexts.

Although it is rarely done, CART should lend itself to model selection methods based on information theory. As all variables are evaluated at each split in a tree, a wide range of alternative trees can be explored (as long as initial variable selection is biologically informed and plausibly tied to hypotheses to avoid data dredging). To date only one tree is typically provided in papers, probably because most software outputs only the *best* tree, but evaluating and reporting alternative trees represents an obvious area of interest. Boosted regression trees (BRT) (Friedman, 2001) and random forests (RF) (Breiman, 2001) represent extensions of CART designed to evaluate multiple trees, and both are being advocated for use in spatial ecology (e.g. Leathwick *et al.*, 2006; Prasad *et al.*, 2006).

SPATIAL PROCESSES EXPLAIN SPATIAL PATTERNS

Legendre (1993) provided a heuristic method for distinguishing environmental and spatial structure in ecological data by means of a partial regression (or constrained ordinations) that partitions '(a) nonspatial environmental variation', '(b) spatially structured environmental variation' and '(c) spatial variation of the target variable(s) that is not shared by the environmental variables' (p. 1666). His use of the language was careful, and this method is now widely used, but it is not uncommon to read that (c) is the *effect of pure space*, or the *effect of spatial processes*. Is it?

Ignoring the problem of effects, consider a very simple scenario. I want to explain statistically the species richness gradient of New World lizards with respect to the current environment. But I do not know much about lizards and use annual precipitation as my measure of environment. I also record the geographical coordinates of my samples/cells to quantify space. I use regression and find, say, that 10% of the variance in richness is related to precipitation in a simple regression and 75% is related to longitude and latitude in a separate multiple regression. I would probably find in a partial regression of the two models that 3–4% of the variation would be attributable to *pure* precipitation (one simple measure of the environment) whereas 50–60% is *pure* space; the rest is overlap and unexplained variation. We may be tempted to conclude that spatial processes are much more important than the environment. But remembering that lizards are extreme solar ectotherms, what if I included annual temperature in my environmental model? How much spatial variation would be independent of the environment now, and what would be the *effect* of *pure space*? The flaw in assigning variation to pure anything is obvious in this example, but clearly, the partial coefficients that underlie all variation partitioning are the result of deciding what variables to include in models and are not interpretable in any pure sense. This is because pattern does not identify process (the second best-known statistical aphorism?). The problem is that few of the processes that operate at each and every scale have been identified and quantified with unambiguous proxy variables for inclusion in geographical studies, so we use spatial proxies instead (latitude, longitude, altitude/elevation, depth, or increasingly, spatial eigenvectors). These can be added to statistical models to quantify the total amount of spatial pattern in the data, but what do they tell us of ecological significance?

Almost all ecological spatial pattern at broad scales is due to direct and indirect responses of organisms to the environment. But the pattern is not the process [see also Landeiro & Magnusson (2011) for discussion of pattern and process in spatial data with respect to conservation biology]. For example, distance of a place from a Pleistocene refugium may pop up in a statistical model accounting for the presence or absence of a plant, but the reason why has to be explained by an interaction of biology and environment; for example, seed size and prevailing wind direction, movement patterns of animal seed-dispersers that are themselves influenced by the environment, the presence of intervening rivers, ocean or mountains and the ability of seeds to get over or around them, the rate of establishment of source populations increasingly nearer the target site as influenced by winter temperatures, summer rainfall and edaphics, and so on. None of these potential influences are *spatial processes*, they are biological processes operating in a spatially structured environment.

Once the biological nature of spatial pattern across broad scales is realized, it becomes apparent that spatial processes *per se* are not common in biogeography, perhaps being largely confined to colonization patterns on islands. Less obvious perhaps is that non-spatial processes are also rare, at least at the spatial scales of interest to biogeographers and macroecologists (although they can be important at smaller scales). That is, at biogeographical scales the distinction between space and environment is usually a false one; the practical difference between them is that we define environment as variables we have measures of and space as variables we do not have measures of. If we had perfect knowledge of any diversity gradient and could measure all relevant influences at all scales (from metres to megametres) *space* would contribute 0% to any statistical model of diversity, there would be zero residual spatial autocorrelation at all scales, and so there would be nothing to correct for or any reason to use spatial regression. The problem is that we cannot measure everything at all scales, and we do not have perfect knowledge, which brings us full circle.

SPATIAL AUTOCORRELATION CAUSES RED SHIFTS IN REGRESSION MODELS

A major reason spatial regression is said to be required for the analysis of spatially structured data is that spatial autocorrelation is thought to overemphasize the *importance* of broad-scale predictors in OLS multiple regression models. This overemphasis is referred to as red shifts [see Lennon (2000) and Diniz-Filho *et al.* (2003) for the origins of the debate]. The key to this is to realize that the red shifts argument has nothing to do with accounting for residual autocorrelation arising from a particular statistical model, but with the belief that the effects of local-scale processes overlap with broad-scale processes across all scales to the extent that a statistical fix is required to correct for them in models dominated by variables operating over broad scales. But it is not that simple. First, the argument about the over-importance of broad-scale predictors in OLS models has it backwards, to the extent that it is an issue at all. As we know, range maps contain false positives, the number of which will be strongly associated with the grain at which the data are binned. Survey data, in contrast, will invariably contain false negatives, and the degree to which this is so is also very sensitive to grain. As species richness estimates are sensitive to grain, sample effort and the length of time over which samples are taken or records are kept, it is a fact that no method exists that can measure accurately a diversity gradient whatever the scale, so there is little point in arguing about precision. But given the data at our disposal, presumably few

workers believe that the spatial autocorrelation found in survey-based data is an artefact, because it reflects the actual presence of individuals in a plot or cell and thus represents biological signal. Data derived from range maps are another story, and the use of range maps represents the source of non-biological spatial structure in geographical ecological data.

Range maps are generated by filling in gaps between occurrence records, which means that the spatial pattern among cells close to each other will tend to be much more distorted than the pattern among distant cells, with the amount of potential distortion increasing with decreasing grain. The effect of this is to obscure small-scale differences in richness, which may cause cells to be even more similar than they really are despite changes in the environment along the gradient. Each case of this adds small amounts of unexplained variation to an OLS model focused on the overall gradient, as reduced variation in richness among nearby cells will partially decouple them from small-scale differences in the broad-scale environmental gradient. The summing of these small amounts of unexplained variation across the data then causes an underestimation of the strength of the broad-scale predictors in OLS models, the opposite of what is claimed [see Hawkins *et al.* (2007) for an empirical example of this phenomenon]. Again, blaming spatial autocorrelation is a mistake, because the same pattern generated by field plots or surveys would contain just as much autocorrelation but no false positives, and hence nothing to correct for. The solution to this data quality problem is to select the smallest grain likely to contain few or no false positives. This is difficult to do in practice, because it is usually impossible to be certain when cells are big enough to eliminate the false positives arising from using range maps, although there has been some work on this issue (Hurlbert & Jetz, 2007; Hawkins *et al.*, 2008). And as we search for the optimal grain to control for data quality we will start to encounter a thorny biological problem as well: all aspects of diversity are scale-dependent, but they do not all respond to changes in scale in the same way, which is one of the reasons partial regression coefficients are sensitive to method. Therefore, the key is not searching for a perfect grain that does not really exist, but in matching the grain of the data with the grain at which predictors (and associated processes) are likely to operate. The following example may seem obvious, but few would expect competitive exclusion to be the dominant driver of a continental richness gradient measured at a grain of 10,000 km^2, whereas post-glacial Pleistocene/Holocene recolonization is unlikely to explain differences between small plots a few kilometres apart. Although the need to be aware of scale sensitivity should be self-evident, it is not unusual to see papers misinterpret regression models because of a mismatch between the presumed process and the grain at which its proxy is measured. A lack of knowledge about how processes influence patterns across scales also underlies the uncertainty associated with the interpretation of spatially explicit regression models.

This brings us to a second issue: what is the cause of coefficient shifts in regression models? It is not spatial autocorrelation, it is multiple regression itself. When one moves between regression methods results can change in complex and unpredictable ways (Bini *et al.*, 2009), which should tell us that the problem lies not in the data, but in the analysis. When we use many of the spatial regression methods, we think we are controlling for some presumed bias caused by not having predictors in our models operating at each and every scale, but in reality we are doing much more than that, changing the effective scale of the analysis, introducing new collinear predictors, modifying covariance structures in non-intuitive ways, and/or making assumptions about how driving variables overlap across scales. Given the many structural problems with all forms of multiple regression and the difficulty in understanding their statistical behaviours when applied to spatially structured ecological data, a claim that one regression method provides stronger inference because it corrects for spatial autocorrelation is not a statement of fact, it is a statement of faith.

Spatial eigenvector analysis is a relatively simple way to incorporate spatial autocorrelation of unknown origin into a model without introducing the additional, hidden effects that may creep in when using many of the spatially explicit regression methods, although like all multiple regression it makes an assumption about the overlap of variables across scales (it assumes there is none, which, if wrong, is at least transparent). Eigenvector analysis based on modelling the spatial structure of residuals allows one to generate a set of scale-specific vectors (often referred to as spatial filters) that can be added to models but which are orthogonal to spatially structured variables already in the model, in which rests the assumption of no across-scale covariance among measured and unmeasured drivers. This then removes all residual spatial autocorrelation without the potential distortion generated by introducing collinearity or scale shifts, and the filters can even capture patterns generated by false positives if there are any. Of course, the filters are not interpretable in an ecological sense, but they do resolve a problem that concerns many workers: that our models do not control for spatial pattern at every scale. But if we are not generating *P*-values from significance tests, what have we really learned of value when compared to simply generating a spatial correlogram that explicitly identifies at what scales a model performs poorly (Diniz-Filho *et al.*, 2003)? Residual autocorrelation is direct evidence that one or more unmeasured spatially structured variables are required to explain all spatial structure in the data, and a correlogram provides important clues with respect to the spatial scale at which these variables are operating. And if we resort to more sophisticated forms of spatial modeling, such as autoregression, the analysis may hide the unexplained variation, but it does not tell us what variables are needed to explain the spatial structure from unknown sources. Nor does attempting to remove autocorrelation using statistical modelling lead to a better understanding of the drivers of geographical patterns if the assumption built into the specific spatial method about how *space* and *environment* overlap across scales is wrong. Indeed, it is the extent to which this assumption is true that underlies arguments about red shifts and whether OLS or spatial regression is *correct*. Unfortunately, this problem will be

difficult to resolve using simulated data given that real data are highly variable in structure and that patterns of scale dependency do not follow simple rules. Interpreting regression models is made even more complicated when the structure of the data violates assumptions of the method. Regression has its uses, but no multiple regression method has logical priority over another given the nature of the data we analyse.

CODA

Legendre (1993) argued that spatial autocorrelation offers new opportunities for understanding pattern in ecological data, but his message has been ignored to the extent that spatial autocorrelation is now considered a bias that must be removed from data rather than embraced. This has led to a confusing literature and a proliferation of increasingly complicated analytical methods that are difficult to evaluate or even understand if you are not a statistician. A tendency to ignore assumptions of many of these complex methods does not help matters. It has also diverted attention away from epistemological and conceptual issues of importance to our field, some of which I have tried to highlight. Although it is self-evident that statistics are an indispensible tool for evaluating data, when we focus too much on methods it is natural to add new layers of complexity as our view becomes narrower and narrower and we try to capture every nuance of our data. But biogeography and geographical ecology are not branches of theoretical statistics, and there does come a point at which analytical complexity begins to interfere with understanding. Geographical analysis can be made quite complicated, but it does not have to be *too* complicated. I have touched on a number of issues, but a major point is about the use of linear multiple regression in biogeography. Let me be as clear as I can: the problem is not whether one form of regression or another is right, it is that linear multiple regression is an inadequate method of analysis given the structure of biogeographical data, and the inherent uncertainty associated with all of its forms is causing disagreements among workers that lead nowhere. But for those who still believe that multiple regression is useful, it comes down to this: do you prefer to use a complicated method for which you cannot evaluate the limitations, or a simple one for which you can?

ACKNOWLEDGEMENTS

I thank L. Mauricio Bini, T. Jonathan Davies, Phillip J. DeVries, J. Alexandre F. Diniz-Filho, Carsten Dormann, Richard Field, Robert P. Freckleton, Pedro R. Peres-Neto, Miguel Á. Rodríguez, Kaustuv Roy, Joseph A. Veech and an anonymous referee for their helpful comments on the manuscript. They do not necessarily endorse everything I have written.

REFERENCES

Bachraty, C., Legendre, P. & Desbruyeres, D. (2009) Biogeographic relationships among deep-sea hydrothermal vent faunas at global scale. *Deep Sea Research Part I: Oceanographic Research Papers*, **56,** 1371–1378.

Beale, C.M., Lennon, J.L., Yearsley, J.M., Brewer, M.J. & Elston, D.A. (2010) Regression analysis of spatial data. *Ecology Letters*, **13,** 246–264.

Bickford, S.A. & Laffan, S.W. (2006) Multi-extent analysis of the relationship between pteridophyte species richness and climate. *Global Ecology and Biogeography*, **15,** 588–601.

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.

Breiman, L. (2001) Random forests. *Machine Learning*, **45,** 5–32.

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1998) *Classification and regression trees*. Chapman and Hall/CRC Press, Boca Raton, FL.

Burnham, K.P. & Anderson, D.R. (2002) *Model selection and multimodel inference: a practical information-theoretical approach*. Springer, New York.

Cassemiro, F.A.S., Barretio, B.S., Rangel, T.F.L.V.B. & Diniz-Filho, J.A.F. (2007) Non-stationarity, diversity gradients and the metabolic theory of ecology. *Global Ecology and Biogeography*, **16,** 820–822.

Cohen, J. (1994) The Earth is round ($p < .05$). *American Psychologist*, **49,** 997–1003.

Cressie, N.A.C. (1993) *Statistics for spatial data*. Wiley, New York.

Currie, D.J. & Paquin, V. (1987) Large-scale biogeographical patterns of species richness of trees. *Nature*, **329,** 326–327.

Diniz-Filho, J.A.F., Bini, L.M. & Hawkins, B.A. (2003) Spatial autocorrelation and red herrings in geographical ecology. *Global Ecology and Biogeography*, **12,** 53–64.

Diniz-Filho, J.A.F., Hawkins, B.A., Bini, L.M., De Marco, P. & Blackburn, T.M. (2007) Are spatial regression methods a panacea or a Pandora's box? A reply to Beale *et al.* (2007) *Ecography*, **30,** 848–851.

Diniz-Filho, J.A.F., Rangel, T.F.L.V.B. & Bini, L.M. (2008) Model selection and information theory in geographical ecology. *Global Ecology and Biogeography*, **17,** 479–488.

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

Fink, D., Hochachka, W.M., Zuckerberg, B., Winkler, D.W., Shaby, B., Munson, M.A., Hooker, G., Riedewald, M., Sheldon, D. & Kelling, S. (2010) Spatiotemporal exploratory models for broad-scale survey data. *Ecological Applications*, **20,** 2131–2147.

Foody, G.M. (2004) Spatial nonstationarity and scale-dependency in the relationship between species richness and environmental determinants for the sub-Saharan endemic avifauna. *Global Ecology and Biogeography*, **13,** 315–320.

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

Fotheringham, A.S., Brunsdon, C. & Charlton, M. (2002) *Geographically weighted regression: the analysis of spatially varying relationships*. Wiley, Chichester.

Freckleton, R.P. (2002) On the misuse of residuals in ecology: regression of residuals vs. multiple regression. *Journal of Animal Ecology*, **71,** 542–545.

Freckleton, R.P. (2009) The seven deadly sins of comparative analysis. *Journal of Evolutionary Biology*, **22,** 1367–1375.

Friedman, J.H. (2001) Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, **29,** 1189–1232.

Galbraith, J.W. & Zinde-Walsh, V. (1997) On some simple autoregression-based estimation and identification techniques for ARMA models. *Biometrika*, **84,** 685–696.

Gittleman, J.L. & Kot, M. (1990) Adaptation: statistics and a null model of estimating phylogenetic effects. *Systematic Zoology*, **39,** 227–241.

Grace, J.B. & Bollen, K.A. (2005) Interpreting the results from multiple regression and structural equation models. *Bulletin of the Ecological Society of America*, **86,** 283–295.

Graham, M.H. (2003) Confronting multicollinearity in ecological multiple regression. *Ecology*, **84,** 2809–2815.

Hawkins, B.A. (2010) Multiregional comparison of the ecological and phylogenetic structure of butterfly species richness gradients. *Journal of Biogeography*, **37,** 647–656.

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.

Hawkins, B.A., Rueda, M. & Rodríguez, M.Á. (2008) What do range maps and surveys tell us about diversity patterns? *Folia Geobotanica*, **43,** 345–355.

Hortal, J., Diniz-Filho, J.A.F., Bini, L.M., Rodríguez, M.Á., Baselga, A., Nogués-Bravo, D., Rangel, T.F., Hawkins, B.A. & Lobo, J.M. (2011) Ice age climate, evolutionary constraints and diversity patterns of European dung beetles. *Ecology Letters*, **14,** 741–748.

Hurlbert, A.H. & Jetz, W. (2007) Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. *Proceedings of the National Academy of Sciences USA*, **104,** 13384–13389.

Landeiro, V.L. & Magnusson, W.E. (2011) The geometry of spatial analyses: implications for conservation biologists. *Natureza & Conservação*, **9,** 7–20.

Leathwick, J.R., Elith, J., Francis, M.P., Hastie, T. & Taylor, P. (2006) Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. *Marine Ecology Progress Series*, **321,** 267–281.

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

Lennon, J.J. (2000) Red-shifts and red herrings in geographical ecology. *Ecography*, **23,** 101–113.

Link, W.A. & Barker, R.J. (2006) Model weights and the foundations of multimodel inference. *Ecology*, **87,** 2626–2635.

Ord, K. (1975) Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, **70,** 120–126.

Peres-Neto, P. (2006) A unified strategy for estimating and controlling spatial, temporal and phylogenetic autocorrelation in ecological models. *Oecologia Brasiliensis*, **10,** 105–119.

Pianka, E.R. & Schall, J.J. (1981) Species densities of Australian vertebrates. *Ecological biogeography of Australia* (ed. by A. Keast), pp. 1676–1694. Dr. W. Junk bv Publishers, The Hague.

Prasad, A.M., Iverson, L.R. & Liaw, A. (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, **9,** 181–199.

Quinn, J.F. & Dunham, A.E. (1983) On hypothesis testing in ecology and evolution. *The American Naturalist*, **122,** 602–617.

Rabinovich, J.E. & Rapoport, E.H. (1975) Geographical variation of diversity in Argentine passerine birds. *Journal of Biogeography*, **2,** 141–157.

Schabenberg, O. & Gotway, C.A. (2005) *Statistical methods for spatial data analysis*. Chapman and Hall/CRC Press, Boca Raton, FL.

Schall, J.J. & Pianka, E.R. (1978) Geographical trends in numbers of species. *Science*, **201,** 679–686.

Shipley, B. (2000) *Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference*. Cambridge University Press, Cambridge.

Tilman, D. & Downing, J.A. (1994) Biodiversity and stability in grasslands. *Nature*, **367,** 363–365.

Tilman, D., Wedin, D. & Knops, J. (1996) Productivity and sustainability influenced by biodiversity in grassland ecosystems. *Nature*, **379,** 718–720.

Tilman, D., Wedin, D. & Knops, J. (1997) The influence of functional diversity and composition on ecosystem processes. *Science*, **277,** 1300–1302.

Tobler, W.R. (1970) A computer movie simulating urban growth in the Detroit region. *Economic Geography*, **46**(Supplement), 234–240.

Whittingham, M.J., Stephens, P.A., Bradbury, R.B. & Freckleton, R.P. (2006) Why do we still use stepwise modelling in ecology and behavior? *Journal of Animal Ecology*, **75,** 1182–1189.

Wilson, P.D. (2008) The pervasive influence of sampling and methodological artefacts on a macroecological pattern: the abundance–occupancy relationship. *Global Ecology and Biogeography*, **17,** 457–464.

BIOSKETCH

Bradford A. Hawkins is interested in biogeography and geographical ecology, with an emphasis on integrating ecological and evolutionary explanations for macroecological patterns. He is not a professional statistician.

Editor: Richard Ladle