Global Ecology and Biogeography, (Global Ecol. Biogeogr.) (2006) **15**, 321–327

ECOLOGICAL SOUNDING

Towards an integrated computational tool for spatial analysis in macroecology and biogeography

Thiago Fernando L. V. B. Rangel*, José Alexandre Felizola Diniz-Filho and Luis Mauricio Bini

Departamento de Biologia Geral, ICB, Universidade Federal de Goiás, CP 131, Goiânia, GO, 74001-970, Brazil

*Correspondence: Thiago Fernando L. V. B. Rangel, Departamento de Biologia Geral, ICB, Universidade Federal de Goiás, CP 131, Goiânia, GO 74001-970, Brazil.
E-mail: t.rangel@terra.com.br

ABSTRACT

Because most macroecological and biodiversity data are spatially autocorrelated, special tools for describing spatial structures and dealing with hypothesis testing are usually required. Unfortunately, most of these methods have not been available in a single statistical package. Consequently, using these tools is still a challenge for most ecologists and biogeographers. In this paper, we present SAM (Spatial Analysis in Macroecology), a new, easy-to-use, freeware package for spatial analysis in macroecology and biogeography. Through an intuitive, fully graphical interface, this package allows the user to describe spatial patterns in variables and provides an explicit spatial framework for standard techniques of regression and correlation. Moran's *I* autocorrelation coefficient can be calculated based on a range of matrices describing spatial relationships, for original variables as well as for residuals of regression models, which can also include filtering components (obtained by standard trend surface analysis or by principal coordinates of neighbour matrices). SAM also offers tools for correcting the number of degrees of freedom when calculating the significance of correlation coefficients. Explicit spatial modelling using several forms of autoregression and generalized least-squares models are also available. We believe this new tool will provide researchers with the basic statistical tools to resolve autocorrelation problems and, simultaneously, to explore spatial components in macroecological and biogeographical data. Although the program was designed primarily for the applications in macroecology and biogeography, most of SAM's statistical tools will be useful for all kinds of surface pattern spatial analysis. The program is freely available at www.ecoevol.ufg.br/sam (permanent URL at http://purl.oclc.org/sam/).

Keywords

Spatial autocorrelation, spatial statistics, modelling, macroecology, biogeography, statistical package, software, SAM.

> 'The process of preparing programs for a digital computer is especially attractive, not only because it can be economically and scientifically rewarding, but also because it can be an aesthetic experience much like composing poetry or music.'
>
> Donald E. Knuth

INTRODUCTION

Ecologists recognize that nearly all macroecological and biodiversity data show strong spatial patterns, driven by spatially structured biological processes, and consequently are often spatially autocorrelated. Following Legendre (1993), spatial autocorrelation may be defined as 'the property of random variables taking values, at pairs of locations a certain distance apart, that are more similar (positive autocorrelation) or less similar (negative autocorrelation) that expected for randomly associated pairs of observations'. The endogenous or exogenous causes of spatial structure or autocorrelation in ecological and biogeographical data are a function not only of how spatially dynamic processes drive observed variables, such as abundance, richness, endemism, body size and range size, but also depend on how the data are collected by spatial sampling schemes (Diniz-Filho *et al*., 2003; Fortin & Dale, 2005).

Spatially autocorrelated data sets present both a potential statistical problem and an opportunity to recognize the importance and understand the causes of the spatial structure in ecology (Legendre, 1993; Diniz-Filho *et al*., 2003). Accordingly, ecologists and biogeographers now dealing with spatially autocorrelated data sets may act in two different ways: (i) they ignore it or brush it aside (jeopardizing publication of their papers in the better

DOI: 10.1111/j.1466-822x.2006.00237.x
www.blackwellpublishing.com/geb

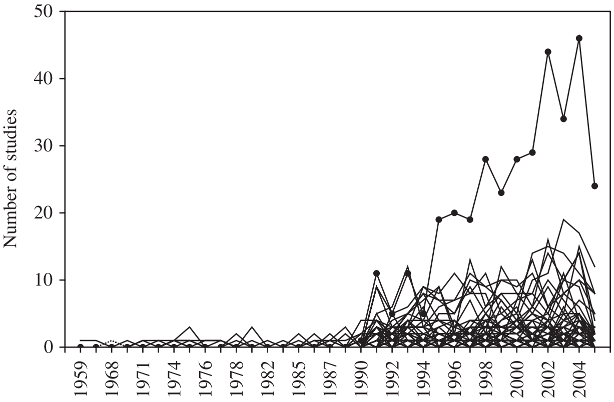

Figure 1 Temporal trends in the number of papers that consider spatial autocorrelation, distributed among 82 subject categories distinguished by ISI Web of Science. Filled circles indicate the number of studies published in ecological journals, whereas other lines indicate trends in all other categories, showing that spatial autocorrelation is now a major issue in ecology and biogeography.

mainstream journals) or (ii) they incorporate realistic assumptions about spatial structure in their analyses and endeavour to understand the underlying spatial processes.

Some idea of the emerging importance of spatial autocorrelation in the worldwide scientific literature can be gleaned using a bibliometric approach. The Thomson Institute (ISI Web of Science) bibliographic data base (1945–20 June 2005) was used to identify all papers indexed that used the terms 'spatial' and 'autocorrelation'. We found a total of 2284 studies distributed in 82 subject categories. After 1990, a clear increase in the number of papers that used the terms 'spatial' and 'autocorrelation' in the title or in the abstract was detected in nearly all subject categories (Fig. 1). This indicates the interest in spatial autocorrelation by different research fields. However, the predominance of papers on spatial autocorrelation published recently in ecological journals is notable, indicating that, at least with respect to this issue, ecologists produce scientific knowledge and not just consume it (Peters, 1991).

Given the increasing importance of spatial autocorrelation analysis in ecology and biogeography, it is important to stress that there are still difficulties in applying these methods routinely, as they are usually not available in most commercial statistical packages. This was one of the principal problems identified by the participants of the workshop on statistical methods for spatial analysis in ecology, sponsored by the National Center for Ecological Analysis and Synthesis (NCEAS). As stated by Liebhold & Gurevitch (2002), 'Ultimately it would be desirable to develop software packages … that are capable of computing a full array of spatial statistics'. However, for most researchers in our field, understanding these complex spatial analyses and modelling approaches is still a challenge. We believe that a key step in improving the understanding of spatial issues in ecology and biogeography is to provide comprehensive, user-friendly and freely available software, together with a useful help file. In this paper, we introduce a computational program named Spatial Analysis in Macroecology (SAM), a new software application for managing spatial analyses that was designed with the particular needs of macroecologists and biogeographers in mind. We also highlight how these analyses interact with some current issues that are of interest for macroecologists and biogeographers.

EXPLORING AND IDENTIFYING THE SPATIAL STRUCTURE OF THE DATA

The first step in every spatial analysis, as in any other statistical procedure, is an exhaustive exploratory data analysis (EDA). Conducting an EDA facilitates immensely the visualization of 'hidden' patterns in a (usually) large-sized macroecological data set (see Rossi *et al.*, 1992 for a discussion of EDA within the framework of spatial analysis). The EDA module of SAM provides five analytical methods to: (i) compute basic and descriptive statistics; (ii) visualize the data (distribution and relationships) by means of graphs (two- and three-dimensional scatterplots, histograms and maps); (iii) create/edit a connectivity matrix using several, alternative criteria (e.g. Delaunay triangulation, Gabriel, Minimum Spanning Tree; Sokal & Oden, 1978a; see more below); (iv) transform the data (log, square-root, etc.); and (v) reduce data dimensionality (principal component analysis). This last approach may be particularly useful when analysing highly correlated variables, such as climatic data at broad scales, reducing the problem of multicolinearity in multiple regression (Philipp, 1993; Graham, 2003). Also, SAM has two special modules that handle presence–absence data in matrices of species X spatial unit (e.g. cells in a 1° × 1° grid), allowing the calculation of richness values using different combinations of species based on macroecological criteria, such as body size, range size, habitat type or evolutionary age [Marquet *et al.*'s (2004) deconstructive approach; see also Cardillo, 2002; Jetz & Rahbek, 2002; Bini *et al.*, 2004; Hawkins *et al.*, 2005; Rahbek, 2005]. For example, a researcher may be interested in calculating spatial pattern in species richness for small-bodied and large-bodied species (Hillebrand & Azovsky, 2001), or computing the average body size of several species in each spatial unit.

SAM includes different modules to test for spatial autocorrelation in a variable (e.g. species richness) and also to distinguish the type of spatial structure (e.g. clines, patches, etc.; see Legendre & Fortin, 1989). Following Rossi *et al.* (1992), these methods (spatial continuity measures) quantify the relationship between the value of a variable in one spatial unit and the value of the same variable in other spatial units.

Because Moran's *I* coefficient is the most commonly used statistic for autocorrelation analyses in macroecological and biogeographical studies (see Badgley & Fox, 2000; Diniz-Filho *et al.*, 2003; and Bini *et al.*, 2004 for recent applications of this coefficient), and because it is statistically robust (Tiefelsdorf, 2000), it is the primary statistic for describing spatial structure in SAM, although semivariance is also available. Moran's *I* can be calculated for multiple distance classes, established using a variety of different criteria, allowing the generation of a graph relating autocorrelation coefficients to the spatial distances among sampling units, called a spatial correlogram (Sokal & Oden, 1978a,b). If there is only one variable in the data set, several ecological issues can be evaluated with a correlogram. Minimally, the analysis of a correlogram furnishes a description of the spatial pattern in the data (Legendre & Fortin, 1989). As we will show below, this is also an important tool for evaluating whether or not a fundamental assumption (independence of residuals) of general linear models holds.

Three issues regarding the description of the spatial structure in the data using autocorrelation coefficients have been discussed inadequately in ecological and biogeographical literature. First, it is possible to estimate Moran's *I* coefficient considering various types of geographical connections (criteria for connecting geographical localities). As indicated by Sokal & Oden (1978a), the criteria for considering a pair of sampling units as connected or not connected depend on the hypothesis being tested (e.g. two sampling units in a stream network, despite the geographical proximity, might be considered unconnected; see Peres-Neto, 2004; Ganio *et al*., 2005). For this reason, SAM offers the opportunity to create an *ad hoc* connectivity matrix that indicates the relationships among sampling units by consideration of the hypothesis under study (e.g. to take into account the presence of an ecological barrier, or to match migration routes or dispersion flows). Generally, connections are treated as some function of geographical proximity. As variations on this approach, SAM offers five standard methods for defining connections among the sampling units: Delaunay Triangulation, Gabriel Criterion, Relative Neighbourhood, Minimum Spanning Tree and Distance Criterion (Legendre & Legendre, 1998; Fortin & Dale, 2005).

Secondly, it is also possible to run a LISA (Local Indicator of Spatial Autocorrelation) analysis (Anselin, 1995; Sokal *et al*., 1998) in SAM, an overlooked method in macroecology. LISA can be used to measure the contribution of each sampling unit to the overall (global) level of spatial autocorrelation (Cocu *et al*., 2005). Finally, it is possible to test the statistical significance of Moran's *I* using randomization (Monte Carlo) (see Manly, 1997; Tiefelsdorf, 2000), which is a reliable way to assess statistical significance, especially for small sample sizes (Sokal & Oden, 1978a,b).

At this point it is worthwhile to mention another very useful feature implemented in SAM. Most spatial analyses use a matrix of distances among pairs of sampling units to describe the spatial relationships in the data, usually assuming a planar surface (Euclidean distance). However, for broad spatial scales (e.g. continents or large domains), which are common in macroecology and biogeography, the calculation of planar distances may bias the spatial relationship among sampling units because of the curvature of the Earth. For this reason, SAM allows the user to compute geodesic surface distance among pairs of sampling points with an accuracy of about 50 m, assuming not only that the Earth is approximately spherical, but also taking into account the actual polar flattening of the Earth and the equatorial bulge. Geodesic surface distances may be used for all spatial analyses in SAM, but only if geographical coordinates are measured in decimal degrees of latitude and longitude.

SAM provides two basic ways to describe and control for spatial structure in the data under the general concept of spatial filtering. The first is the well-known trend surface analysis (TSA), which has been extensively discussed elsewhere (Wartenberg, 1985; Davis, 1986). SAM allows the automatic calculation of TSA polynomial expansions up to the 6th order. However, a highly recommended alternative is Principal Coordinates of Neighbour Matrices (PCNM; see Borcard & Legendre, 2002; Borcard *et al*., 2004; Diniz-Filho & Bini, 2005) because of the efficiency of this method in partitioning variation between spatial and environmental components (Borcard *et al*., 2004). Further, the importance of scale in detecting the magnitude and direction of relationships among variables is well known, and PCNM can deal effectively with this issue (Whittaker *et al*., 2001; Rahbek, 2005).

Spatial filters obtained by TSA or PCNM can be used in different ways, depending on how spatial patterns are taken into account. One approach is to use them to remove all spatial structure from the data and work only with (residual) non-spatial components to evaluate, for example, the effect of predictors on richness. In this example, applying this approach would be appropriate if broad-scale spatial processes did not contain, or could not reliably reveal, information regarding causal process associated with richness, due to the confounding effects of intrinsic and extrinsic processes affecting this variable. Alternatively, these filters can be treated as candidate explanatory variables together with other, environmental predictors. With this approach, the effects of environmental predictors are evaluated as partial effects, taking space into account explicitly (see below). These two different approaches may produce different results, depending on the collinearity between predictors and space.

A different subject in the analysis of spatial pattern is the identification of patches and regions in space (Fortin & Dale, 2005). This approach may be especially necessary in a broad or multiple scale spatial analysis, when several ecological processes may be driving independently different regions of an observed spatial pattern. However, the likelihood of detecting these ecological processes depends on our ability to delineate ecological patches, boundaries, edges or ecotones (Oden *et al*., 1993; Fortin, 1994; Fortin & Drapeau, 1995; Fortin *et al*., 2000), and also on the ecological processes under investigation, the sampling design and the employed analytical methodology. To help macroecologists and biogeographers detect edges, two edge detection algorithms, called triangulation-wombling and lattice-wombling (Fortin & Drapeau, 1995; Fagan *et al*., 2003), are available in SAM. The difference between these two methods concerns rules to join adjacent sampling units although, in both methods, a region with spatial discontinuity is detected by the steepness and direction of the slope of the plane formed by a set of joint sampling units (for details, see Fortin & Dale, 2005).

MODELLING AND HYPOTHESIS TESTING

The most frequently discussed issue in the ecological literature regarding spatial autocorrelation is the inflation of Type I errors in significance tests of correlation and regression analyses (Legendre, 1993; Diniz-Filho *et al*., 2003; and references therein). If two (or more) variables are each strongly spatially autocorrelated, spatial units close in geographical space are partially redundant with respect to the information they provide about the relationships between variables. In other words, in the presence of spatial autocorrelation, the number of degrees of freedom is overestimated and, consequently, confidence intervals are much narrower than they should be. This may cause an error in judging the statistical significance under a null hypothesis. Thus, the non-independence of data caused by spatial structure can lead these analyses to be liberal and, thus, even variables that are in fact correlated weakly will appear to yield significant coefficients due to the confounding effects of space. Significance inflation is important when trying to understand the effect of different predictors on a response variable, such as which environmental factors give better explanations for spatial variation in species richness. Such analyses are beyond the simple description of spatial variation described previously, and taking the spatial dimension into account usually improves the ability to model the spatial variation and understand the causal factors underlying it.

The simplest solution for testing a correlation coefficient in the presence of autocorrelation is to adjust the number of degrees of freedom, an approach developed by Clifford *et al.* (1989) and Dutilleul (1993) (see Legendre *et al.*, 2002 and Hawkins *et al.*, 2005 for discussions and applications). SAM provides two estimators to calculate the geographically effective number of degrees of freedom, both using spatial correlograms of the raw variables to be correlated. Although they usually provide similar results, they have different computational requirements (because Dutilleul's approach is more computationally intensive, it will demand more time for large matrices).

For a simple spatial modelling, SAM provides tools for ordinary least squares (OLS) regression, with three special features: (i) the evaluation and mapping of spatial structure in model residuals (which may reveal the need for explicit spatial modelling); (ii) partial regression analysis, using up to a 6th order polynomial expansion of geographical coordinates; and (iii) the calculation of the Akaike information criterion (AIC) (Burnham & Anderson, 2002; Johnson & Omland, 2004), allowing an easy and powerful comparative evaluation of model fit when competing hypothesis are confronted with data. AIC as computed by SAM is based on the sum of squares of residuals and provides a approximation of AIC based on likelihood under a normal distribution of error terms (Mangel & Hilborne, 1997).

When strong autocorrelation is found in model residuals, alternative modelling strategies are available. The first is to include filters (TSA or PCNM), as discussed previously, along with the predictive variables, to minimize residual autocorrelation (see Diniz-Filho & Bini, 2005). However, SAM also allows the fitting of explicit spatial regression models to data. Three forms of autoregression models (ARM) are available.

The first set of routines allows estimation of 'lagged-models' (see Haining, 1990, 2002), which are based on fitting a pure autoregressive model that describes the spatial structure of only response variable **Y**, given by:

$$\mathbf{Y} = \rho\mathbf{WY} + \mathbf{e}$$

where ρ is the autoregression parameter, and the matrix **W** contains neighbour weights (w_{ij}), indicating the relationships among spatial units. The elements w_{ij} can be given by the connectivity matrices discussed previously or as an inverse power function of geographical distances (d_{ij}), given by functions of the form $w_{ij} = d_{ij}^{\alpha}$, where α is an additional parameter that regulates the relationship and that usually improves the performance of the model (Davis, 1986).

Because ARM allows a description of the spatial structure in data, it could just as well have been offered within the 'Structure' section of SAM. It appears, instead, in the 'Modelling' section as the basis of more complex spatial regression models that can be used to evaluate the effects of predictors on the response variable, by adding additional terms (see Haining, 1990, 2002). The first option for added terms assumes that the autoregressive process modelled by ARM occurs only in the response variable (lagged-response model), and thus includes a term for the spatial autocorrelation in **Y**, as in ARM (above), but also includes the standard term for the predictors in OLS. The model then becomes:

$$\mathbf{Y} = \rho\mathbf{WY} + \mathbf{X}\beta + \mathbf{e}$$

where β is a vector representing the slopes associated with the predictors in the original predictor matrix **X**. Alternatively, autocorrelation can affect both response and predictor variables (lagged-predictor model, or Durbin econometric model — see Anselin, 1988). In this case yet another term must appear in the model, which now takes the form:

$$\mathbf{Y} = \rho\mathbf{WY} + \mathbf{X}\beta + \mathbf{WX}\gamma + \mathbf{e}$$

where γ represents the autoregressive parameters of each of the predictors. Note that in this more complex model there is an autoregression parameter for each predictor. In all cases, AIC and residual spatial autocorrelation can be used to choose among alternative models, which can also be generated using different α values, as discussed for the simple ARM. These last two models are in fact fitted by working with the residuals e of the pure ARM, described above, in a standard OLS regression. For this reason, these models can also be interpreted as techniques that filter the effect of space (see Haining, 1990; Anselin, 2002). For the lagged-response model, the ARM residuals of the response variable are regressed against the original predictors using OLS, whereas in the lagged-predictor model the ARM residuals of the response variable is regressed against ARM residuals of each predictor variable (Haining, 1990, p. 347).

Another explicit way to take autocorrelation into account in a regression is by changing the estimator of the vector of slopes (β) by applying a generalized least-squares (GLS) model that incorporates spatial structure directly into model residuals (see Selmi & Boulinier, 2001; Hawkins & Diniz-Filho, 2002; Evans *et al.*, 2005). This vector is given by:

$$\beta = (\mathbf{X}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}\,\mathbf{C}^{-1}\mathbf{Y}$$

where **C** is the covariance among residuals. In the standard OLS model, it is assumed that $\mathbf{C} = \mathbf{I}\sigma^2$ (**I** is an identity matrix, and σ^2 is the variance of the residuals), so that residuals are independent. However, it is possible to model the values of **C** using a semi-variogram, fitted by exponential, powered-exponential, Gaussian, spherical, hole-effect, Matérn, truncated-linear or pure-nugget models (Legendre & Legendre, 1998; Banerjee *et al.*, 2004). Fitting these models usually requires an iterative process, and SAM allows the user to manually fit the semi-variograms and decide among them based on visual inspection or on their explanatory ability. After fitting the semi-variograms, regression slopes of GLS are obtained – in effect, a two-step evaluation (modelling OLS residuals then using their spatial structure to approximate GLS estimation). This approach is sometimes called 'kriging regression' (Cressie, 1993).

The GLS framework also allows the direct computation of other forms of spatial regressions that have been used recently in ecology and macroecology (e.g. Lichstein *et al.*, 2002; Dark, 2004; Tognelli & Kelt, 2004), such as Simultaneous and Conditional Autoregression Models (SAR and CAR, respectively), by computing the matrix **C** with different formats (see Haining, 1990; Cressie, 1993). For SAR, the covariance matrix among residuals is given by:

$$\mathbf{C} = \sigma^2 \, [(\mathbf{I} - \rho\mathbf{W})^{\mathrm{T}}]^{-1}[\mathbf{I} - \rho\mathbf{W}]^{-1}$$

where σ^2 is the variance of the residuals and **I** is an $n \times n$ identity matrix. For CAR, this matrix is given as:

$$\mathbf{C} = [(\sigma^2 \, \mathbf{W}_{i+})\mathbf{I}](\mathbf{I} - \rho\mathbf{W})^{-1}.$$

SAM also computes the moving average (MA) model, where the covariance matrix among residuals is given by:

$$\mathbf{C} = \sigma^2 \, [(\mathbf{I} + \rho\mathbf{W})(\mathbf{I} + \rho\mathbf{W})].$$

See Wall (2004) for the conditions under which these three models can be expressed, as described above, as functions only of autoregressive parameter ρ and neighbour weight matrix **W**. Note that, in these models, the matrix of neighbour weights (**W**) may also be computed as an inverse power function of geographical distances among sampling units ($w_{ij} = d_{ij}^{\alpha}$).

The GLS is then a regression in which the spatial component is explicitly modelled in the residual terms, defined by the fitted semi-variogram. Thus, these residuals contain a strong spatial component, which must be decomposed into spatially structured residuals and a pure error term using Cholesky decomposition (see Cresci, 1992, p. 202; Haining, 2002, p. 333). This error vector **e**, or noise component, is given as:

$$\mathbf{e} = \mathbf{L}^{-1} \, (\mathbf{Y} - \mathbf{X}\beta)$$

where β is the vector of estimated slopes and $\mathbf{LL}^{\mathrm{T}} = \mathbf{C}$, so that the **L** matrix can be obtained by the Cholesky decomposition of the covariance among residuals. In the GLS, this error term is then a function of the model used to fit the semi-variogram, whereas in SAR, CAR and MA models the error term is a function of the autoregressive parameter according to the functions defined above. The effectiveness of GLS-based models, in terms of taking autocorrelation into account, can be judged by the absence of spatial structures in this error term. These properties can be visualized in SAM by spatial correlograms of residual and error terms in the GLS-based routines.

The r^2 due to explanatory variables for all these GLS-based spatial models is obtained using Nagelkerke (1991) general formulation for coefficients of determination, given as:

$$r^2 = 1 - e^{-2/n(l_{\mathrm{A}} - l_0)}$$

where n is number of spatial units, l_{A} is the log-likelihood of the model, and l_0 is log-likelihood of the null model fitted with only the intercept (Lichstein *et al.*, 2002). The 'full r^2' of the model, which incorporates the joint effects of the spatial component and the predictors, is given as the complement of the squared linear correlation coefficient between response and the error term, $r^2 = 1 - \text{Pearson}\,[\mathbf{Y}; \mathbf{e}]^2$.

A full comparison of these various spatial regression models in macroecology and biogeography is still lacking, primarily with respect to understanding how choosing each of them will affect the relative importance of predictors driving biodiversity at different spatial scales (see Diniz-Filho *et al.*, 2003; Tognelli & Kelt, 2004; Ferrer-Castan & Vetaas, 2005). This is not just a statistical issue, because the choice among different predictors may provide support for alternative, and sometimes competing, biogeographical hypothesis (Hawkins *et al.*, 2003; Currie *et al.*, 2004). Hopefully, this new software will help researchers to use more spatial models and to perform more accurate hypothesis testing and data exploration and, thus, allow a deeper understanding of complex ecological and biogeographical patterns and processes.

COMPUTATIONAL FEATURES AND SOFTWARE AVAILABILITY

SAM is able to open input files in any of three data formats: Excel, dBASE and ASCII (tab-delimited). Analytical results from all analyses appear in text windows in SAM modules and may be exported by Windows's copy-paste tool. Any new variables created by SAM during analysis (e.g. regression residuals, principal components, spatial filters, etc.) may be saved in the original data file or exported from SAM to another file in any permissible format.

In principle, SAM has no computational limits, but practical limits are set by the computational power of the machine that runs it. The constraint on working with very large data sets (above 3000 cases) is probably not on opening the matrix, but on the time and memory needed to perform some of the analyses, which require varying amounts of memory and CPU time. The time needed to compute the more computationally demanding statistics in SAM depends on the computer, the data set size and the operation of interest. Our experience shows that running most currently available PCs (about 2.5 GHz, 256 Mb of RAM), SAM can easily handle matrices of up to 1000 cases for all routines (albeit with variable computer times). The Dell Precision 450 Workstation (3.5 GHz, 4 GB RAM) on which SAM was developed was able to run easily approximately 5000 cases for all routines.

SAM is freeware, and the present version is less than 10 MB. Researchers interested in using SAM may download it from the official web site: www.ecoevol.ufg.br/sam (permanent URL at http://purl.oclc.org/sam/). Three complete, real sample data sets are distributed together with the SAM application file (birds of South America, birds and mammals of the Brazilian cerrado, birds and mammals of the western hemisphere), and these data sets may be helpful for new users as trial examples. Prior to downloading SAM, users are required to provide their name, institution, country and e-mail address, so that a user list can be compiled to determine how many researchers are interested in and using the package, to justify the grant used to build the software and to allow us to contact users about future versions.

ACKNOWLEDGEMENTS

We thank (alphabetically) Miguel Araújo, Daniel Borcard, Rob Colwell, Nick Gotelli, Brad Hawkins, Holger Kreft, Matt Talluto and about 20 initial users in two SAM courses (Universidade Federal de Goiás, Brazil, and Universidad de Alcalá, Spain), and 70 other users since the first SAM release (June 2005), for suggestions on the software. Rob Colwell and Brad Hawkins also kindly revised and suggested improvements on the text. We are especially grateful to Daniel Borcard and Ole Vetaas for the insightful comments and suggestions on the initial version of the manuscript. SAM was developed on a Dell Precision 450 Series Workstation provided by CNPq/SECTEC-GO (PRONEX project no. 23234156). The authors received graduate and productivity fellowships from CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico).

REFERENCES

Anselin, L. (1988) *Spatial econometrics: methods and models*. Kluwer Academic Publishers, Boston.

Anselin, L. (1995) Local indicators of spatial association — LISA. *Geographical Analysis*, **27**, 93–115.

Anselin, L. (2002) Under the hood. Issues in the specification and interpretation of spatial regression models. *Agricultural Economics*, **27**, 247–267.

Badgley, C. & Fox, D.L. (2000) Ecological biogeography of North American mammals: species density and ecological structure in relation to environmental gradients. *Journal of Biogeography*, **27**, 1437–1467.

Banerjee, S., Carlin, B.P. & Gelfand, A.E. (2004) *Hierarchical modeling and analysis for spatial data*. Chapman & Hall/CRC, Boca Raton.

Bini, L.M., Diniz-Filho, J.A.F. & Hawkins, B.A. (2004) Macroecological explanations for differences in species richness gradients: a canonical analysis of South American birds. *Journal of Biogeography*, **31**, 1819–1827.

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

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

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

Cardillo, M. (2002) Body size and latitudinal gradients in regional diversity of New World birds. *Global Ecology and Biogeography*, **11**, 59–65.

Clifford, P., Richardson, S. & Hemon, D. (1989) Assessing the significance of the correlation between 2 spatial processes. *Biometrics*, **45**, 123–134.

Cocu, N., Harrington, R., Hulle, M. & Rounsevell, M.D.A. (2005) Spatial autocorrelation as a tool for identifying the geographical patterns of aphid annual abundance. *Agricultural and Forest Entomology*, **7**, 31–43.

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

Currie, D.J., Mittelbach, G.C., Cornell, H.V., Field, R., Guégan, J.-F., Hawkins, B.A., Kaufman, D.M., Kerr, J.T., Oberdorff, T., O'Brien, E.M. & Turner, J.R. (2004) Predictions and tests of climate-based hypotheses of broad-scale variation in taxonomic richness. *Ecology Letters*, **7**, 1121–1134.

Dark, S.J. (2004) The biogeography of invasive alien plants in California: an application of GIS and spatial regression analysis. *Diversity and Distributions*, **10**, 1–9.

Davis, J.C. (1986) *Statistics and data analysis in geology*, 2nd edn. Wiley, New York.

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

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.

Dutilleul, P. (1993) Modifying the *t*-test for assessing the correlation between 2 spatial processes. *Biometrics*, **49**, 305–314.

Evans, K.L., Warren, P.H. & Gaston, K.J. (2005) Species-energy relationship at macroecological scale: a review of the mechanisms. *Biological Reviews*, **80**, 1–25.

Fagan, W.F., Fortin, M.-J. & Soycan, C. (2003) Integrating edge detection and dynamic modeling in quantitative analysis of ecological boundaries. *Bioscience*, **53**, 730–738.

Ferrer-Castan, D. & Vetaas, O.R. (2005) Pteridophyte richness, climate and topography in the Iberian Peninsula: comparing spatial and nonspatial models of richness patterns. *Global Ecology and Biogeography*, **14**, 155–165.

Fortin, M.-J. (1994) Edge-detection algorithms for 2-dimensional ecological data. *Ecology*, **75**, 956–965.

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

Fortin, M.-J. & Drapeau, P. (1995) Delineation of ecological boundaries — comparison of approaches and significance tests. *Oikos*, **72**, 323–332.

Fortin, M.-J., Olson, R.J., Ferson, S., Iverson, L., Hunsaker, C., Edwards, G., Levine, D., Butera, K. & Klemas, V. (2000) Issues related to the detection of boundaries. *Landscape Ecology*, **15**, 453–466.

Ganio, L.M., Torgersen, C.E. & Gresswell, R.E. (2005) A geostatistical approach for describing patterns in stream networks. *Frontiers in Ecology and the Environment*, **3**, 138–144.

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

Haining, R. (1990) *Spatial data analysis in the social and environmental sciences*. Cambridge University Press, Cambridge.

Haining, R. (2002) *Spatial data analysis*. Cambridge University Press, Cambridge.

Hawkins, B.A. & Diniz-Filho, J.A.F. (2002) The mid-domain effect cannot explain the diversity gradient of Nearctic birds. *Global Ecology and Biogeography*, **11**, 419–426.

Hawkins, B.A., Diniz-Filho, J.A.F. & Soeller, S.A. (2005) Water links the historical and contemporary components of the Australian bird diversity gradient. *Journal of Biogeography*, **32**, 1035–1042.

Hawkins, B.A., Porter, E.E. & Diniz-Filho, J.A.F. (2003) Productivity and history as predictors of the latitudinal diversity gradient of terrestrial birds. *Ecology*, **84**, 1608–1623.

Hillebrand, H. & Azovsky, A.I. (2001) Body size determines the strength of the latitudinal diversity gradient. *Ecography*, **24**, 251–256.

Jetz, W. & Rahbek, C. (2002) Geographic range size and determinants of avian species richness. *Science*, **297**, 1548–1551.

Johnson, J.B. & Omland, K.S. (2004) Model selection in ecology and evolution. *Trends in Ecology and Evolution*, **19**, 101–108.

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

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

Legendre, P. & Fortin, M.J. (1989) Spatial pattern and ecological analysis. *Vegetatio*, **80**, 107–138.

Legendre, P. & Legendre, L. (1998) *Numerical ecology*. Elsevier, Amsterdam.

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

Liebhold, A.M. & Gurevitch, J. (2002) Integrating the statistical analysis of spatial data in ecology. *Ecography*, **25**, 553–557.

Mangel, R. & Hilborne, M. (1997) *The ecological detective: confronting models with data.* Princeton University Press, Princeton.

Manly, B.F.J. (1997) *Randomization, bootstrap and Monte Carlo methods in biology*. Chapman & Hall, London.

Marquet, P.A., Fernández, M., Navarrete, S.A. & Valdovinos, C. (2004) Diversity emerging: toward a deconstruction of biodiversity patterns. *Frontiers of biogeography: new directions in the geography of nature* (ed. by M.V. Lomolino & L.R. Heaney), pp. 191–210. Sinauer, Sunderland.

Nagelkerke, N.J.D. (1991) A note on a general definition of the coefficient of determination. *Biometrika*, **78**, 691–692.

Oden, N.L., Sokal, R.R., Fortin, M.J. & Goebl, H. (1993) Categorical wobbling — detecting regions of significant change in spatially located categorical variables. *Geographical Analysis*, **25**, 315–336.

Peres-Neto, P.R. (2004) Patterns in the co-occurrence of fish species in streams: the role of site suitability, morphology and phylogeny versus species interactions. *Oecologia*, **140**, 352–360.

Peters, R.H. (1991) *A critique for ecology*. Cambridge University Press, Cambridge.

Philipp, T.E. (1993) Multiple regression: herbivory. *Design and analysis of ecological experiments* (ed. by S.M. Scheiner and J. Gurevitch), pp. 183–210. Chapman & Hall, New York.

Rahbek, C. (2005) The role of spatial scale and the perception of large-scale species-richness patterns. *Ecology Letters*, **8**, 224–239.

Rossi, R.E., Mulla, D.J., Journel, A.G. & Franz, E.H. (1992) Geostatistical tools for modeling and interpreting ecological spatial dependence. *Ecological Monographs*, **62**, 277–314.

Selmi, S. & Boulinier, T. (2001) Ecological biogeography of Southern Ocean islands: the importance of considering spatial issues. *The American Naturalist*, **158**, 426–437.

Sokal, R.R. & Oden, N.L. (1978a) Spatial autocorrelation in biology. 1. Methodology. *Biological Journal of the Linnean Society*, **10**, 199–228.

Sokal, R.R. & Oden, N.L. (1978b) Spatial autocorrelation in biology. 2. Some biological implications and four applications of evolutionary and ecological interest. *Biological Journal of the Linnean Society*, **10**, 229–249.

Sokal, R.R., Oden, N.L. & Thomson, B.A. (1998) Local spatial autocorrelation in biological variables. *Biological Journal of the Linnean Society*, **65**, 41–62.

Tiefelsdorf, M. (2000) *Modelling spatial processes: the identification and analysis of spatial relationships in regression residuals by means of Morans' I.* Cambridge University Press, Cambridge.

Tognelli, M.F. & Kelt, D.A. (2004) Analysis of determinants of mammalian species richness in South America using spatial autoregressive models. *Ecography*, **27**, 427–436.

Wall, M.M. (2004) A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference*, **121**, 311–324.

Wartenberg, D. (1985) Canonical trend–surface analysis — a method for describing geographic patterns. *Systematic Zoology*, **34**, 259–279.

Whittaker, R.J., Willis, K.J. & Field, R. (2001) Scale and species richness: towards a general, hierarchical theory of species diversity. *Journal of Biogeography*, **28**, 453–470.

BIOSKETCHES

Thiago Fernando L. V. B. Rangel is interested in statistical and computational methods applied to macroecology and evolutionary biology. Current projects involve the development and implementation of spatial simulation models to understand the role of evolutionary- and climate-based factors that drive latitudinal gradients in species richness and macroecological patterns. He also leads the long-term project of SAM design and development.

José Alexandre F. Diniz-Filho is interested in statistical methods applied to macroecology and conservation biology. Current projects involve application of spatial autocorrelation analysis and phylogenetic comparative methods to understand ecological processes associated with gradients in species richness and the application of spatial statistics to conservation reserve design and to establish conservation priorities.

Luis Mauricio Bini is interested in statistical methods applied to biodiversity analyses and limnology. Current projects involve the analysis of spatial population synchrony of aquatic assemblages in reservoirs and floodplains. He is also interested in how population dynamics is linked with more general biodiversity patterns, mainly the relationship between species diversity and ecosystem stability.

Editor: David Currie