00000100000
00001010000
00010001000
00100000100
01000000010
10000000001
Software notes

**Ecography 33: 46–50, 2010**
doi: 10.1111/j.1600-0587.2009.06299.x

Subject Editor: Carsten Rahbek. Accepted 15 December 2009

# SAM: a comprehensive application for Spatial Analysis in Macroecology

**Thiago F. Rangel, Jose Alexandre F. Diniz-Filho and Luis Mauricio Bini**

*T. F. Rangel (rangel.sam@gmail.com), Dept of Ecology and Evolutionary Biology, Univ. of Connecticut, Storrs, CT 06269-3043 USA. – J. A. F. Diniz-Filho and L. M. Bini, Depto de Ecologia, Univ. Federal de Goiás, Goiânia, Goiás, Brasil.*

SAM (Spatial Analysis in Macroecology) is a freeware application that offers a comprehensive array of spatial statistical methods, focused primarily on surface pattern spatial analysis. SAM is a compact, but powerful stand-alone software, with a user-friendly, menu-driven graphical interface. The methods available in SAM are the most commonly used in macroecology and geographical ecology, and range from simple tools for exploratory graphical analysis (e.g. mapping and graphing) and descriptive statistics of spatial patterns (e.g. autocorrelation metrics), to advanced spatial regression models (e.g. autoregression and eigenvector filtering). Download of the software, along with the user manual, can be downloaded online at the SAM website: <www.ecoevol.ufg.br> (permanent URL at <http://purl.oclc.org/sam/>).

Today there are many software applications and packages available for spatial statistical analysis. Some of them are stand-alone applications that offer several methods (e.g. Passage <www.passagesoftware.net>, GeoDa <geodacenter.asu.edu>), while others are specific to particular methods (e.g. GWR3 <http://ncg.nuim.ie/ncg/GWR/software.htm>, SpaceMaker2 <www.bio.umontreal.ca/casgrain/en/labo/spacemaker.html>, ModTTest <www.bio.umontreal.ca/legendre/indexEn.html>) or collections of routines within a general purpose statistical platform (e.g. SpDep for R, EconoTools for MatLab). SAM (Spatial Analysis in Macroecology, Rangel et al. 2006) is a compact, but powerful stand-alone freeware application, compiled for the MS Windows environment, with a user-friendly, menu-driven graphical interface. SAM offers a comprehensive array of spatial statistical methods. The methods available in SAM are the most commonly used in macroecology and geographical ecology, ranging from simple tools for exploratory graphical analysis (e.g. mapping and graphing) and descriptive statistics of spatial patterns (e.g. autocorrelation metrics), to advanced spatial regression models (e.g. autoregression and eigenvector filtering).

Since SAM's first release, in August 2005, it has been downloaded about 9300 times (Fig. 1a), by researchers working in >60 countries around the world. By tracking scientific publications that cite the original SAM paper (Rangel et al. 2006, Fig. 1b), we identified 165 studies that cited SAM, of which 83% reported that SAM was directly used for spatial statistical analysis. Among those studies, 77% used SAM to investigate general ecological questions, whereas 25% used for biodiversity conservation, 22% for physical geography and 9% for questions related to evolutionary biology. These papers were collectively published in 45 different journals, by authors from 33 different countries. The most commonly used methods implemented in SAM were Moran's I correlogram (43%), Dutilleul's (1993) estimator of effective sample size used in correlation analysis (12%), spatial auto-regression models (SAR, CAR or GLS, 11%) and spatial eigenvector mapping (7%).

SAM has been under continuous development and expansion (Table 1). SAM now uses extremely optimized linear algebra libraries for the most computer-intensive methods, so that time-consuming procedures (e.g. involving eigenanalysis) are now must faster. Here we show how the most important features currently available in SAM evolved, while highlighting the new and improved features available in SAM v4, released in March 2010.

The data table in SAM is a rectangular matrix of numeric values, in which columns are variables and rows are individual observations (e.g. grid cells), formatted in tab-delimited text (ASCII) (*.txt or *.sam), dBase (*.dbf), MS Excel (*.xls) or ESRI shapefile (*.shp and companion files). Geographic coordinates must be included as two of the columns (variables) in the data file. In addition to the main data table, recent versions of SAM also allow the input of species presence/absence matrices, in which each species is represented in its own column, while rows are locations in which the species is present (1) or absent (0). Presence/absence matrices can be used, for instance, to compute richness patterns considering different criteria (e.g. body size and taxonomic structures; Bini et al. 2004,

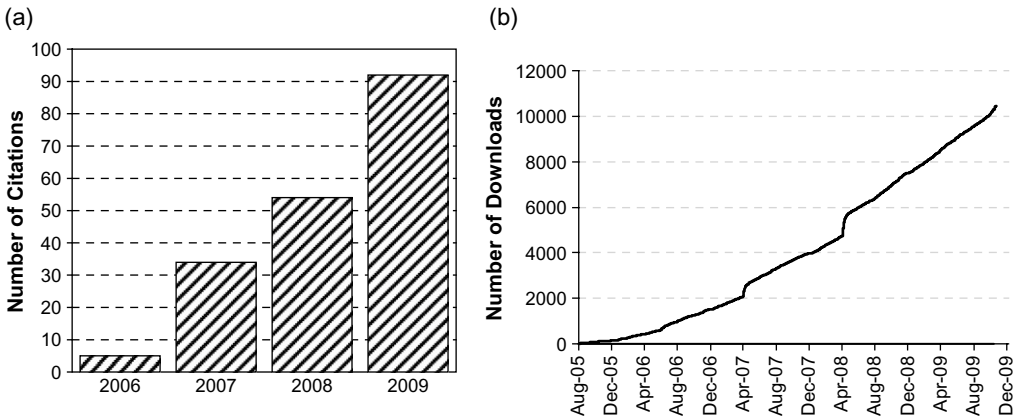


Figure 1. (a) Time-series of number of scientific publications that cite the original SAM paper (Rangel et al. 2006). (b) Time series of cumulative number of SAM downloads since first release (August 2005). Discontinuities in April 2007 and April 2008 were caused by intensified download activity following the releases of SAM v2 and v3.

Terribile et al. 2009). In addition, if a matrix of species' traits (with species in rows and species' traits in columns) is available, then individual species in the presence/absence matrix can be selected according to a given trait (e.g. species with body size larger than the average body size), or species traits can be mapped in the geographical space, given the species assemblage in each location. This is a very useful tool for those interested in some of the most frequent investigated macroecological patterns, as for instance, the Bergmann's and Rapoport's rules.

Previous SAM versions were mostly dedicated to data analysis, and most of data processing relevant to macroecological studies had to be done with the aid of a GIS software. The GIS environment implemented in the current SAM version, however, allows users to easily prepare data for macroecological analysis without any additional software. Grids can be generated in any resolution and extent, using equal area square or hexagonal cells, and they can be saved into shapefiles. Also, because shapefiles have become a standard format to share information on species distributions (range polygons or points), SAM can process the distribution of each species to record its presence or absence in each grid cell, and thus generate presence/absence matrices directly from shapefiles. Finally, from ESRI rasters or text files, environmental layers can be downscaled to the resolution of the grid by calculating mean and standard deviation of all observations within each grid cell, which then become additional variables in the main data matrix.

The graphical exploratory data analysis (GEDA) is one of the most important steps in statistical analysis (Tukey 1980). For this reason, one of SAM's greatest strengths is its rich collection of graphical analytical tools and the simplicity of using and editing them. All charts, which may be drawn with just a few clicks, allow zooming, scrolling and changing colors, maximizing investigators' capacity to find patterns and identify particular details in the data. Colors are abundantly used to highlight patterns

Table 1. The evolution of the most used modules available in SAM. Greek letters denote the versions of the modules (α: first; β: second; γ: third; δ: forth). GEDA stands for Graphical Exploratory Data Analysis, PAM stands for Presence/Absence Matrix and SEVM stands for Spatial Eigenvector Mapping).

| | SAM v1 Aug-05 | SAM v2 Aug-06 | SAM v3 Aug-08 | SAM v4 Mar-10 |
|---|---|---|---|---|
| GEDA tools | α | β | γ | δ |
| Moran's I and Auto-Correlogram | α | α | α | β |
| Spatial Correlation | α | α | α | β |
| Regression and Partial Regression | α | β | β | γ |
| PAM and Spp. Attributes Mapping | | α | α | β |
| Principal Component Analysis | | α | α | α |
| Auto-Regression: Lagged | | α | α | β |
| Auto-Regression: SAR/CAR | | α | α | β |
| Auto-Regression: GLS | | α | α | β |
| SEVM | | | α | β |
| Model Selection and Multi-Model Inference | | | α | α |
| Logistic Regression | | | α | β |
| Geographically Weighted Regression | | | α | α |
| GIS Processing and Mapping | | | α | β |
| Pattern Finder | | | α | β |
| Ripley's K | | | | α |
| Join-Count Analysis | | | | α |
| Mantel Test | | | | α |
| ANOVA | | | | α |

in the data or to superimpose multiple plots in the same panel. For example, in two-dimensional scatter-plots, polynomial regression lines can be easily drawn to highlight the relationship between two variables. In the three-dimensional scatter-plot, tilting and rotating allow that visual inspection of the data can be done from any perspective. In the new version, the residuals plot applies a well designed set of color gradients in a scatter plot and in a map simultaneously. This tool graphically displays the correlation between two variables and is ideal for evaluating the geographic structure in model's residuals, as it uses different color gradients to differentiate under- and over-estimated values.

Maps are the most important exploratory tools in spatial data analysis, which is why they are fully embedded in each of SAM's analytical modules. SAM allows users to easily draw one or multiple maps simultaneously, which facilitates the visual comparison of the spatial patterns in different variables. The map module in SAM allows re-sizing, zooming and scrolling simple maps, as well as inspecting values by moving the cursor over the map. An even more advanced mapping module is enabled when the main data is extracted from an ESRI shapefile. For example, one may overlay multiple map layers, which may be regular or irregular polygons, and points, to produce publication-quality maps. The graduated color gradients, with customizable classes, are applicable to each shapefile, with automatically generated legends.

The strength of the relationships among variables can change across space (see GWR below). For example, water availability is thought to affect species richness in the tropics, whereas temperature is the most important driver of species richness at higher latitudes (Hawkins et al. 2003). Pattern Finder is a new tool available in SAM that graphically links scatter plots, maps and tables to aid the identification of geographically structured relationships. Using this tool, one can select cells in a map, then the points in the scatter plot and the rows in a spreadsheet that refer to the selected cells are highlighted. The selection of the data may also be made directly from the scatter plot or the spreadsheet. This is an especially useful tool to detect outliers or mistyping.

One of the most important steps in exploratory analysis of spatial data is to measure the magnitude and direction of spatial autocorrelation, which has been 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) than expected for randomly associated pairs of observations'' (Legendre 1993). Moran's I coefficient is one of the most commonly used descriptors of spatial autocorrelation. Moran's I can be calculated for individual distance classes (e.g. from 0 to 300 km, 300 to 600 km), producing a plot known as a spatial correlogram. Besides a standard spatial correlogram, SAM's current version also computes asymmetric correlograms, directional correlograms (Rosenberg 2000), Anselin's Moran's I scatter plot (Anselin 1996), and local Moran's I (LISA, Sokal et al. 1998). Also, a new module in SAM implements join-count analysis, which measures the magnitude of spatial autocorrelation in binary data, and is thus very useful to describe the spatial pattern in the distribution of species. Still in the context of spatial autocorrelation, a new important feature in SAM 4.0 is that autocorrelation can be evaluated in multidimensional data using Mantel test (Manly 1998). This technique is widely used in ecology and evolutionary biology to evaluate if the (dis)similarity among samples (many metrics are available) is structured in geographic space. One of the advanced features on SAM implementation of Mantel test is the ability to perform a Mantel correlogram, which separates the geographic space into sequential distance classes to aid the identification of changes in the strength of correlation between matrices (e.g. compositional similarity against a distance matrix) at different scales.

The problem of inflated type I error rates and model instability that may arise from violation of the assumption of residuals independence in ecological models are now well-known (Legendre 1993, Schabenberger and Gotway 2005, Diniz-Filho et al. 2008, Cliff and Ord 2009). In SAM, this assumption can be easily checked for by evaluating the spatial correlogram of regression residuals that is automatically calculated when the regressed data is spatially explicit. However, researchers have been gradually abandoning classical null-hypothesis testing when the actual goal of the analysis is to confront multiple competing hypotheses (Hilborne and Mangel 1997, Burnham and Anderson 2002). Instead, they have been adopting the information theoretic approach to select the best model among a large set of competing models, or combining the most parsimonious models as a function of their rank. SAM performs model selection and multi-model inference employing the Akaike information criterion (AIC), which provides a parsimonious balance between model predictive power and complexity. Thus, when a set of competing explanatory variables are defined by the researcher, SAM evaluates models that emerge from all possible combinations of individual variables, and ranks them according to their AIC value and derived statistics (e.g. Akaike's weights and delta AIC). In addition, when the goal is to estimate model parameters or to generate a single predictive model, a ''multi-model'' consensus is calculated by averaging and weighting the estimated model parameters as function of Akaike weights. Although this module is based on a standard OLS approach, spatial structure may be easily incorporated by adding spatial covariates as ''fixed'' predictors in the model selection procedure (Diniz-Filho et al. 2008).

When a matrix of explanatory variables represents two or more sets of competing hypotheses, it is possible to quantify the explanatory power due to individual sets of variables as well as the magnitude of redundancy between the sets. Partial regression analysis has been widely applied in spatial ecology to quantify how the total variation in a response variable can be attributed to the independent effects of the 1) environmental variation not structured in space, 2) spatially structured environmental variation, 3) intrinsic spatially contagious processes, and the 4) unexplained variation. The partial regression module in the current version SAM allows users to define up to three sets of variables, which could be, for example, contemporary environmental factors (e.g. temperature), historical factors (e.g. mean root distance of a phylogenetic

tree) and spatial covariates (e.g. polynomial expansions of geographic coordinates).

A strategy commonly employed to account for spatial autocorrelation in regression analysis is to explicitly incorporate in the model the spatial relationship between pairs of sites. The family of statistical techniques that employ this strategy is collectively known as autoregression, or spatial regression models (Dormann et al. 2007), because they require the estimation of the autoregressive parameter to measure the magnitude of autocorrelation in the data. There are several autoregressive (AR) models available in SAM, including: pure (PAR), lagged-response (LRAR), lagged-predictor (LPAR), simultaneous (SAR), conditional (CAR) and moving-average (MAAR) autoregression. In addition, researchers may also use a semi-variogram to define a variance-covariance matrix and incorporate the spatial structure in a Generalized Least Squares (GLS) model (a technique known as kriging regression).

Among the techniques available today for spatial regression, one of the most flexible and statistically powerful is spatial eigenvector mapping (SEVM, Borcard et al. 2004, Diniz-Filho and Bini 2005, Griffith and Peres-Neto 2006, Bini et al. 2009). SEVM comes in various flavors, depending on how the matrix of spatial relationships among pairs of observations is defined. The module implemented in SAM has been continuously improved, and the current version has three important new features: 1) allows both binary connectivity and continuous distance matrices, 2) provides additional ways to select eigenvectors, including the minimization of Moran's I in model residuals, and 3) both explanatory variables (e.g. environmental factors) and spatial eigenvectors can be analyzed simultaneously within the SEVM module, which enables the automated computation of a partial regression analysis between explanatory variables and spatial predictors.

Both spatial and non-spatial regression models actually require careful evaluation of the stationarity assumption (a lack of importance in the absolute geographical position for estimating model parameters), as violations to this assumption may lead to biases in estimated parameters (Fotheringham et al. 2002). For example, if the direction and magnitude of an ecological process shift from one region to another, the parameter estimated by a global stationary model weights the strength and direction of the process in both regions, which may lead to the conclusion that the processes is globally irrelevant to the observed pattern. Thus, the Geographically Weighted Regression (GWR), now implemented in SAM, is an important method because it allows users to evaluate possible violations of the stationarity assumption and to estimate geographically varying model parameters that may then be biologically interpretable (Cassemiro et al. 2007).

In modeling process, another new possibility in the recent version of SAM is to use presence-absence data to model species' distributions, in the context of niche modeling or species distribution modeling (SDM) (Elith et al. 2006). Although SAM is not particularly designed to run the many different algorithms available for SDM, it now provides a routine for logistic regression that can be used for SDM when presence and absence data are available. This tool can be coupled with other richness analyses and allows a first evaluation of species' environmental drivers and their relationship with other macroecological patterns (Terribile et al. 2009). Moreover, spatial autologistic model is also available in SAM. This model uses the information on the relative position of the species occurrence to generate a spatial weighting covariate, and aims to improve the model predictive power by accounting for stochastic processes driving species distribution, such as species' dispersal capacity (Segurado et al. 2006). Dormann et al. (2007) recently used artificial simulation to show that autologistic models sometimes underestimate the effect of environmental factors, although his analyses have been questioned by Betts et al. (2009).

Download of the software, along with the user manual, can be found online at the SAM website: <www.ecoevol.ufg.br/sam> (<http://purl.oclc.org/sam/>).

To cite SAM or acknowledge its use, cite this Software note as follows, substituting the version of the application that you used for "Version 4":

Rangel, T. F., Diniz-Filho, J. A. F. and Bini, L. M. 2010. SAM: a comprehensive application for Spatial Analysis in Macroecology. – Ecography 33: 46–50, (Version 4).

*Acknowledgements* – We thank SAM users for the continuous stimulus on the SAM project. We are also grateful to all users that provided feedback on the software, especially bug reporting. TFR is funded by CAPES/Fulbright fellowship, Univ. of Connecticut and National Science Foundation (DEB-0639979 and DBI-0851245). JAFD-F and LMB have been continuously funded by CNPq research fellowships.

## References

Anselin, L. 1996. The Moran scatterplot as an ESDA tool to assess local instability in spatial association. – In: Fischer, M. et al. (eds), Spatial analytical perspectives in GIS. Taylor and Francis, pp. 111–125.

Betts, G. M. et al. 2009. Comments on "Methods to account for spatial autocorrelation in the analysis of species distributional data: a review". – Ecography 32: 374–378.

Bini, L. M. et al. 2004. Macroecological explanations for differences in species richness gradients: a canonical analysis of South American birds. – J. Biogeogr. 31: 1819–1827.

Bini, L. M. et al. 2009. Coefficient shifts in geographical ecology: an empirical evaluation of spatial and non-spatial regression. – Ecography 32: 193–204.

Borcard, D. et al. 2004. Dissecting the spatial structure of ecological data at multiple spatial scales. – Ecology 85: 1826–1832.

Burnham, K. P. and Anderson, D. R. 2002. Model selection and multimodel inference: a practical information-theoretic approach, 2nd ed. – Springer.

Cassemiro, F. A. S. et al. 2007. Non-stationarity, diversity gradients and the metabolic theory of ecology. – Global Ecol. Biogeogr. 16: 820–822.

Cliff, A. D. and Ord, J. K. 2009. What were we thinking? – Geogr. Anal. 41: 351–363.

Diniz-Filho, J. A. F. and Bini, L. M. 2005. Modelling geographical patterns in species richness using eigenvector-based spatial filters. – Global Ecol. Biogeogr. 14: 177–185.

Diniz-Filho, J. A. F. et al. 2008. Model selection and information theory in geographical ecology. – Global Ecol. Biogeogr. 17: 479–488.

Dormann, C. F. et al. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. – Ecography 30: 609–628.
Dutilleul, P. 1993. Modifying the t-test for assessing the correlation between two spatial processes. – Biometrics 49: 305–314.
Elith, J. et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. – Ecography 29: 129–151.
Fotheringham, A. S. et al. 2002. Geographically weighted regression: the analysis of spatially varying relationships. – Wiley.
Griffith, D. A. and Peres-Neto, P. R. 2006. Spatial modeling in ecology: the flexibility of eigenfunction spatial analysis. – Ecology 87: 2603–2613.
Hawkins, B. A. et al. 2003. Energy, water, and broad-scale geographic patterns of species richness. – Ecology 84: 3105–3117.
Hilborne, R. and Mangel, M. 1997. The ecological detective: confronting models with data. – Princeton Univ. Press.
Legendre, P. 1993. Spatial autocorrelation: trouble or new paradigm? – Ecology 74: 1659–1673.
Manly, B. F. J. 1998. Randomization, bootstrap and Monte Carlo methods in biology. – Chapman and Hall.
Rangel, T. F. L. V. B. et al. 2006. Towards an integrated computational tool for spatial analysis in macroecology and biogeography. – Global Ecol. Biogeogr. 15: 321–327.
Rosenberg, M. S. 2000. The Bearing correlogram: a new method of analyzing directional spatial autocorrelation. – Geogr. Anal. 32: 267–278.
Schabenberger, O. and Gotway, C. A. 2005. Statistical methods for spatial data analysis. – Chapman and Hall/CRC.
Segurado, P. et al. 2006. Consequences of spatial autocorrelation for niche-based models. – J. Anim. Ecol. 43: 433–444.
Sokal, R. R. et al. 1998. Local spatial autocorrelation in biological variables. – Biol. J. Linn. Soc. 65: 41–62.
Terribile, L. C. et al. 2009. Richness patterns, species distribution and the principle of extreme deconstruction. – Global Ecol. Biogeogr. 18: 123–136.
Tukey, J. W. 1980. We need both exploratory and confirmatory. – Am. Stat. 34: 23–25.