Using simulated data to explore the effects of spatial structure, sampling strategy, and statistical methods on species distribution models

Jennifer A. Miller

Department of Geography and the Environment
The University of Texas at Austin
1 University Station - A3100
Austin, TX 78712-1098
Telephone: (001-512-471-5116)
Email: Jennifer.miller@austin.utexas.edu

1. Introduction

The ability to map and monitor animal and plant species distributions has become more important in the context of awareness of environmental change and its effects on biodiversity (Franklin 1995; Guisan and Zimmermann 2000; Guisan and Thuiller 2005). Species distribution models (SDM) are based on the quantification of species-environment relationships and have seen increasing use as the availability of geospatial technologies and spatial analysis tools expands. SDM require digital maps of important environmental variables, such as topography and climate, as well as spatial information on the species attribute of interest (e.g. presence/absence, type, abundance), usually from a sample of locations. Although used quite widely in ecological applications with often very sophisticated statistical techniques, most SDM are still developed without considering the spatial autocorrelation that is inherent in biogeographical data (Miller et al 2007). More traditional statistical methods used in SDM, such as generalized linear models (GLM) are typically based on the implicit assumption that the distribution of species is random and, therefore, each observation is independent. This assumption violates one of the basic tenets of geography, the direct relationship between distance and likeness, as well as of ecological theory, that elements of an ecosystem close to one another are more likely to be influenced by the same generating process and will therefore be similar. Ignoring spatial autocorrelation in biogeographical data can lead to poorly specified models in general and inflated significance estimates for predictor variables in particular (Legendre 1993).

Some of the spatial structure in species distributions can be explained by the predictor variables used in the model. Environmental variables such as precipitation, temperature and elevation exhibit spatial autocorrelation, some of which is responsible for spatial clustering in species distribution, but remaining spatial autocorrelation can result from unmeasured or misspecified environmental variables or biotic processes that cause spatial clustering, such as competition, dispersal, predation. The majority of previous SDM studies ignore spatial autocorrelation altogether, and those that do acknowledge it usually consider it a nuisance and attempt to manipulate the sampling scheme to avoid autocorrelated observations. The potential predictive ability of spatial autocorrelation is only recently being explored in SDM and similar research (Augustin et al 1998; Ferrier et al 2002; Miller, 2005; Osborne et al, 2007).

The above-mentioned studies typically compare one method that incorporates spatial autocorrelation to a method does not (eg. autologistic to logistic regression), and use compiled species data often sampled for a different objective. Species data that have been collected for other purposes are often based on a previous sampling paradigm intended to reduce autocorrelation in samples by, for example, increasing distance between samples. When spatial model results are compared to nonspatial model results, it is difficult to untangle the effects of actual spatial structure in the data, sampling method, and modelling method. In addition, using a single spatial model to determine whether including spatial dependence is preferable to not including it may be too simplistic, as different spatial models should perform differently as a function of the spatial structure in the data.

This work focuses on an extensive comparison among different statistical methods that incorporate spatial autocorrelation, using different sampling strategies and intensities, and simulated binary (presence/absence) species data with varying levels of spatial autocorrelation. I focus on simulated data here so that the spatial autocorrelation can be adjusted in a ‘known’ way. Additionally, by using simulated data, the confounding effects associated with using data sampled for different purposes can be minimized.

The response variable used here is the distribution of *Yucca brevifolia* (Joshua Tree), the indicator species for the Mojave Desert. Complete data representing the distribution of Joshua Tree species is simulated for a study area comprised of about 60,000 30m grid cells based on quantified relationships between environmental variables (topographical and climatic) from published studies (Miller and Franklin, 2002; Miller, 2005), as well as a factor representing the strength of autocorrelation.

The level of spatial autocorrelation in the species data consists of: none (random), weak, medium, and strong. Different sampling strategies and intensities for collecting the data to be used in the models are also tested. The sampling strategies are: random, gradient-directed sampling (see Franklin et al, 2000), and a hierarchical systematic method that forces pairs of observations at two different spatial lags. Samples are collected from the simulated distribution at the following intensities (proportion of sample): 0.1, 0.01, and 0.005. As the response data used here are binary, the non-spatial model to which the spatial models are compared is logistic regression (LR). Two different spatial models are compared to the nonspatial model. Autologistic regression (AL) is the most similar to LR, the difference being that it includes an additional term that describes neighbourhood values; and geographically weighted regression (GWR), a method that allows model coefficients to vary spatially.

Model results are compared to the remaining (‚known’) grid cell values, and accuracy is assessed using receiver-operator characteristic curves (ROC), a threshold- and prevalence-independent measure used with binary data (see Fielding and Bell, 1997).

2. Acknowledgements

Funding and support for this research was provided by the National Science Foundation (award #0451486).

3. References

Augustin, N., Mugglestone, M., and Buckland, S.. (1998). The role of simulation in modelling spatially correlated data. *Environmetrics* 9: 175-196.

Ferrier, S., Watson, G., Pearce, J., and Drielsma, M. (2002). Extended statistical approaches to modelling spatial pattern in biodiversity in north-east New South Wales. I. Species level modelling. *Biodiversity and Conservation* 11: 2275-2307.

Fielding, A., and Bell, J. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation* 24: 38-49

Franklin, J. (1995) Predictive vegetation mapping: geographic modeling of biospatial patterns in relation to environmental gradients. *Progress in Physical Geography* 19: 474-499.

Franklin, J., Keeler-Wolf T, Thomas K, Shaari D, Stine P, Michaelsen J, and Miller J (2001) Stratified sampling for field survey of environmental gradients in the Mojave Desert Ecoregion. In: Millington A, Walsh S, and Osborne P, eds. *GIS and Remote Sensing Applications in Biogeography and Remote Sensing*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 229-251.

Guisan, A., and Zimmermann, N. (2000). On the use of static distribution models in Ecology. *Ecological Modelling* 135: 147-186.

Guisan, A., and Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology Letters* 8: 993-1009.

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

Miller, J., and Franklin, J. (2002). Modeling the distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence. *Ecological Modelling* 157: 227-247.

Miller, J. (2005) Incorporating spatial dependence in predictive vegetation models: Residual interpolation methods. *The Professional Geographer* 57(2): 169-184.

Miller, J., Franklin, J., and Aspinall, R. (2007). Incorporating spatial dependence in predictive vegetation models. *Ecological Modelling* 202(3-4):225-242.

Osborne, P., Foody, G., and S. Suárez-Seoane. (2007). Non-stationarity and local approaches to modelling the distributions of wildlife. *Diversity and Distributions* 13:313-323.