Spatial autocorrelation in species distribution models: autologistic model with covariates vs. BioMod

Mejía-Domínguez, N.R.[1,*], Díaz-Ávalos, C.[1] and Ochoa-Ochoa, L.M.[2]

[1]*Departamento de Probabilidad y Estadística, Universidad Nacional Autónoma de México,Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas; nancy@sigma.iimas.unam.mx; zhangkalo@gmail.com*
[2]*Center of Macroecology, Evolution and Climate, University of Copenhagen, Denmark; Leticia.Ochoa@snm.ku.dk*
[*] *nancy@sigma.iimas.unam.mx*

Abstract. *Several methods has been proposed to modeling species distributions, also a few models consider explicitly spatial autocorrelation or spatial neighborhood. We compare a model based on autologistic model with covariates with standard methods frequently used in species distribution models. Accuracy assessment is important in species distribution models. The value of a model depends on the accuracy of its outputs. We evaluated the accuracy in terms of ROC analysis and the* F*-measure proposed by Li and Gou 2013 for virtual species and for real species data.*

Keywords. *Besag's model; Covariates; ROC, virtual species*

---

1 Introduction

Species distribution models have become a dominant paradigm for quantifying species-environment relationships, these models simplify the complexity of these relationships and in consequence the models have an evident uncertainty degree. In the last two decades, has been proposed a wide range of algorithms to model species distribution that differs in terms of accuracy and of distributional performance, size, form and continuity of the resultant area (Elith *et al*., 2006; Dorman *et al*., 2007). Additionally, often the only species data available are the geographical coordinates of sites where the species was observed - so-called presence-only data - as recorded by observers (Aarts 2012; Dormann *et al*., 2012; Liu *et al*., 2013).

Owing to the binary nature of the data, using the fitted models, species occurrence probability is predicted and can be projected onto a map of the region. The estimation of probability of presence or occurrence (*pi*) may be based on logistic regression, although other approaches use climatic envelopes, decision trees, and Markov random fields, etc (). In all these models, the probabilities are converted to

presence-absence areas using a binary classification (usually the user choose a threshold value pi). This procedure involves that wrong classification occurs in two forms: because the species is classified as missing in a area where it is present (false negative) or because an area where the species is missing is classified as occupied by it (false positive) (Golicher *et al*., 2012). With a virtual species we know a true distribution and therefore it is possible evaluate the quality of maps produce by the different algorithms using ROC analysis (Egan 1975; Mason and Graham 2002; Hoeting *et al*., 2001). Li and Guo (2013) developed two new accuracy metrics for presence-only data.

Modeling species distributions is complicated by a phenomenon known as spatial autocorrelation, this phenomenon occurs when the values of variables sampled at nearby locations are not independent. However, spatial autocorrelation may be seen as both an opportunity and a challenge for spatial analysis. It is an opportunity when it provides useful information for inference of process from pattern. Neighboring values are expected to be similar to the focal value for ecological reasons such as dispersal, which will lead to higher abundance of off- spring near to the parent organism, or aggregative behavior, leading, e.g., to colonial breeding (Danchin *et al*., 2004). The autologistic regression calculates an extra explanatory variable that captures the effect of other response values in the spatial neighborhood. However, is necessary extended the model to an autologistic model with covariates, as propose Augustin *et al*. 1996. The autologistic model with covariates assumes site dependence within neighborhoods and uses covariates to improve prediction. Diaz-Avalos (2007) proposed a model to construct maps of probability of presence for biological species, based only on records of presence. The model is based on ideas proposed by Heikkinen and Hogmander (1994) and Hoeting *et al*. (2001).

In this study, we use a model proposed by Diaz-Avalos (2007) and compare with models included in BIOMOD (a computer platform for ensemble forecasting of species distributions) with ROC analysis and F-measure proposed by Li and Gou 2013. The models included in BIOMOD are: Generalised linear models, Generalised additive models, Classification tree analysis, Artificial Neural Networks, Surface range envelope, Generalised bosting model, and Multiple adaptive regression splines (Thuiller 2003).

2 Methods

In order to evaluate the performance of the autologistic model with covariates and models included in BIOMOD, we constructed an artificial distribution map for a hypothetical species and for a frog Mexican species. *Craugastor rugulosus* is a species of frog in the Craugastoridae family its natural habitats are subtropical or tropical dry forests, subtropical or tropical moist montane forests, and rivers. In both cases, the accuracy of differents models was evaluated in terms of ROC and F-measure.

2.1 Hierarchical Spatial Model

Let $\mathbf{x} \in \{0,1\}^N$, denote the unknown true presence-absence map of the target species and let $\mathbf{y} \in \{0,1\}^k$ denote the observations. $y_i = 1$ implies that the species is present in pixel i and in this particular case we will assume that x_i=1. The value $y_i = 0$ is attached to pixels for which the value of x_i is unknown, either because the pixel was visited and the species was not detected (because the species was absent or because it is too shy), or because the pixel was not visited, this is,

As mentioned before, there is a chance that the species is not detected so it may happen that $y_i = 0|x_i = 1$,

$$y_i = \begin{cases} 1 & \text{If the species has been recorded at pixel } i \\ 0 & \text{otherwise} \end{cases}$$

and we will denote by g, $P[Y_i = 0|x_i = 1]$. The density of the observations $y_i = 1$ is a function of x_i and g, i.e.

$$f(y_i|x_i, g) = g^{x_i(1-y_i)}(1-g^{x_i})^{y_i}$$

Assuming that the observations y_i are conditionally independent given x and g, the likelihood is

$$L(\mathbf{x}, g; \mathbf{y}) = \prod_{i=1}^{n} g^{x_i(1-y_i)}(1-g^{x_i})^{y_i} \tag{1}$$

For the first step, our interest is the estimation of the probability of presence, using the data $y = (y_1, ..., y_k)$.

We will consider making inferences on x from its conditional distribution given y using

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{x}, \mathbf{y}) = p(x)L(\mathbf{x}, g; \mathbf{y})$$

where $p(x)$ is the prior distribution of the true but unknown presence-absence map. Note that inferences on x could be done, in principle, using maximum likelihood by maximizing equation (1) over g and x, but such maximization is awkward because the high number of possible configurations for x. The likelihood for the autologistic model is analytically intractable, except in trivial cases, so alternative estimation methods are necessary as pseudolikelihood approaches for parameter estimation. Wu and Huffer (1997) use a Markov chain Monte Carlo (MCMC) to approximate the maximum likelihood estimates of the parameters. Bayesian approaches to estimate the parameters of the basic autologistic model have been developed by several authors (Heikkinen and Hogmander 1994). We used the method proposed by Diaz-Avalos (2007), e.i. parameter estimates were obtained via Gibbs sampler by sampling from the full conditionals, except for the conjugate distributions, in all the other cases sampling from the full conditionals was done using the Metropolis-Hastings algorithm (Gilks *et al.*, 1996).

Acknowledgments. Work partially funded by grant post-doctoral fellowship from the Universidad Nacional Autónoma de México (UNAM- México)

References

[1] Aarts, G. (2012). Comparative interpretation of count, presence-absence and point methods for species distribution models. *Methods in Ecology and Evolution*, **3**: 177-187.

[2] Adler, F.R. (1996). A model of self-thinning through local physiological models competition. *Proc. Natl. Acad. Sci. USA* **93**: 9980-9984.

[3] Augustin, N.H., Mugglestone, M.A. and Buckland, S.T. (1996) An autologistic model for the spatial distribution of wildlife. *Journal of Applied Ecology*, **33**: 339-347.

[4] Besag, J.E. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society Series B*, **40**: 147-174.

[5] Diaz-Avalos, C. (2007). Spatial Modeling of Habitat Preferences of Biological Species using Markov Random Fields. *Journal of Applied Statistics*, **34**: 807-821

[6] Dormann, C., M. McPherson, J., B. Araujo, M., Bivand, R., Bolliger, J., Carl, G., G. Davies, R., Hirzel, A., Jetz, W., Daniel Kissling, W., Kuhn, I., Ohlemuller, R., R. Peres-Neto, P., Reineking, B., Schroder, B., M. Schurr, F. and Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, **30**: 609-628.

[7] Dormann, C.F., Schymanski, S. J., Cabral, J., Chuine, I., Graham, C., Hartig, F., Singer, A. (2012). Correlation and process in species distribution models: bridging a dichotomy. *Journal of Biogeography*, **39**: 2119-2131.

[8] Egan, J.P. (1975) Signal Detection Theory and ROC Analysis . *New York: Academic Press*.

[9] Elith J., Graham C.H., ND Anderson R.P. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**:129-151.

[10] Gilks,W.R., Richardson, S. and Spiegehalter, D.J. (1996) *Markov Chain Monte Carlo in Practice*. NewYork, Chapman and Hall.

[11] Golicher, D., Ford, A., Cayuela, L., and Newton, A. (2012). Pseudo-absences, pseudo-models and pseudo-niches: pitfalls of model selection based on the area under the curve. *International Journal of Geographical Information Science*, **26**: 2049Ð2063.

[12] Heikkinen, J. and Hogmander, H. (1994) Fully Bayesian approach to image restoration with an application to biogeography. *Applied Statistics*, **43**:569-582.

[13] Hoeting, J. Leecaster, M. and Bowden, D. (2001) An improved model for spatially correlated binary responses. *Journal of Agricultural Biological and Ecological Statistics*, **5**:102-114.

[14] Liu, C., White, M., and Newell, G. (2013). Selecting thresholds for the prediction of species occurrence with presence-only data. *Journal of Biogeography*, **40**:778-789.

[15] Li, W., and Guo, Q. (2013). How to assess the prediction accuracy of species presence-absence models without absence data? *Ecography*, **36**: 788-799.

[16] Mason, S., and Graham, N. (2002). Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves?: Statistical significance and interpretation. *Quarterly Journal of the Royal* 2145-2166.

[17] Wu, H., and Huffer,F . W. (1997), Modelling the Distribution of Plant Species Using the Autologistic Regression Model. *Environmental and Ecological Statistics*,**4**:49-64.