Revista Mexicana de Biodiversidad 85: 189-199, 2014
DOI: 10.7550/rmb.36723

Potential species distribution modeling and the use of principal component analysis as predictor variables

Modelado de la distribución potencial de especies y el uso del análisis de componentes principales como variables predictoras

Gustavo Cruz-Cárdenas[1], Lauro López-Mata[2]✉, José Luis Villaseñor[3] and Enrique Ortiz[3]

[1]Centro Interdisciplinario de Investigación para el Desarrollo Integral Regional-Instituto Politécnico Nacional-Michoacán, COFAA. Justo Sierra 28, 59510 Jiquilpan, Michoacán, Mexico.
[2]Posgrado en Botánica, Colegio de Postgraduados. Km. 36.5 Carretera Federal México-Texcoco. Montecillo, 56230 Texcoco, Estado de México, Mexico.
[3]Departamento de Botánica, Instituto de Biología, Universidad Nacional Autónoma de México. Apartado postal 70-367, 04510 México, D. F., Mexico.
✉ *laurolopezmata@gmail.com; lauro@colpos.mx*

Abstract. Prior to modeling the potential distribution of a species it is recommended to carry out analyses to reduce errors in the model, especially those caused by the spatial autocorrelation of presence data or the multi-collinearity of the environmental predictors used. This paper proposes statistical methods to solve drawbacks frequently disregarded when such models are built. We use spatial records of 3 species characteristic of the Mexican humid mountain forest and 2 sets of original variables. The selection of presence-only records with no autocorrelation was made by applying both randomness and pattern analyses. Through principal component analysis (PCA) the 2 sets of original variables were transformed into 4 different sets to produce the species distribution models with the modeling application in Maxent. Model precision was higher than 90% applying a binomial test and was always higher than 0.9 with the area under the curve (AUC) and with the partial receiver operating characteristic (ROC). The results show that the records selected with the randomness method proposed here and the use of the PCA to select the environmental predictors generated more parsimonious predictive models, with a precision higher than 95%, and in addition, the response variables show no spatial autocorrelation.

Key words: randomness test, pattern analysis, spatial autocorrelation.

Resumen. Cuando se modela la distribución potencial de una especie es deseable efectuar algunos análisis previos para reducir errores en el modelo resultante, especialmente los ocasionados por la autocorrelación espacial de los registros de presencia y la correlación entre los predictores ambientales utilizados. En este trabajo se proponen métodos estadísticos que sirven para resolver estos inconvenientes que con frecuencia se presentan al elaborar los modelos de distribución potencial. Se emplearon los registros de presencia de 3 especies características del bosque húmedo de montaña de México y 2 conjuntos de variables originales. A los datos de presencia se les aplicó un análisis de aleatoriedad y de patrones para seleccionar registros no autocorrelacionados. Mediante análisis de componentes principales (PCA), los 2 conjuntos de variables originales se transformaron en 4 conjuntos distintos para generar los modelos de distribución de especies utilizando el algoritmo Maxent. La precisión de los modelos fue mayor al 90% con una prueba binomial y mayor de 0.9 del área bajo la curva (AUC) con la característica operativa del receptor parcial (ROC). Los resultados muestran que la selección de registros por el método de aleatoriedad propuesto y el uso de componentes principales como predictores ambientales generan modelos predictivos más parsimoniosos, con una precisión mayor al 95%, además de que sus variables predictivas no presentan autocorrelación espacial.

Palabras clave: prueba de aleatoriedad, análisis de patrón, autocorrelación espacial.

Introduction

The models that predict the potential distribution of species through the combination of presence-only records and digital layers of environmental variables are of great interest in both theoretical and applied disciplines (Guisan and Thuiller, 2005; Elith and Leathwick, 2009a; Peterson et al., 2011). Such models use the association between environmental variables, presumably of predictive value, and the species occurrence records; thus are identified

Recibido: 02 abril 2013; aceptado: 07 octubre 2013

the environmental conditions where a species could survive indefinitely (Pulliam, 2000; Guisan and Thuiller, 2005; Elith and Leathwick, 2009b). This approach is especially important to produce basic information for such disciplines as biogeography, conservation biology, ecology, evolutionary biology, and others (Guisan and Zimmermann, 2000; Guisan and Thuiller, 2005; Elith and Leathwick, 2009b; Peterson et al., 2011).

Species distribution models implicitly suppose that the geographical data points for species records are independent, although this is not necessarily true. In addition, the environmental layers used as hypothetical predictive variables and associated to the geographical records of species also show problems of spatial autocorrelation. The spatial autocorrelation is the degree of dependency of variables in geographical space (Cressie, 1991; Legendre, 1993; Anselin et al., 2004); accordingly, disparity among variable values is strongly influenced by the distances among geographical data points where a species has been observed (Anselin et al., 2004; Segurado et al., 2006). Spatial autocorrelation represents an intrinsic characteristic in most of the geospatial data (Legendre, 1993; Segurado et al., 2006) and it can be an important bias in most geospatial analyses (Anselin et al., 2004). Spatial autocorrelation inflates type I errors of traditional statistics and it can affect the estimated parameters in model selection (Lennon, 2002).

The species distribution models obtained from a large data set of associated environmental covariates often inherently result in multi-collinearity, a statistical problem defined as a high degree of correlation among covariates. Multi-collinearity is a serious statistical problem in non-experimental situations, where the researcher has no control of the risk associated to hypothetical factors related to independent variables. Multi-collinearity is found, for instance, when many covariates are used as predictor variables to model selection and several of them measure similar phenomena. This is so because in most cases the researcher does not have *a priori* knowledge on which predictive environmental variables should be included in the model. However, the researcher must have a model in mind that usually includes a large number of predictive variables and hopes that using an appropriate statistical analysis will provide him/her with a correct model. It should be taken into account that multi-collinearity does not violate the assumptions that underlie to the statistical analysis, i.e., its presence does not affect the estimate of the dependent variable. In other words, estimation values for the dependent variable are the best unbiased estimates from the conditional population average. However, the existence of multi-collinearity tends to inflate both the variances of predicted values of the response variable and the variances of the estimated parameters. Therefore, if one considers that multi-collinearity is present in a dataset, it is important to know how the linear relationships are among the predictive environmental variables. For these reasons, it is critical both to the researcher as to the research to be sure that those environmental predictive variables are orthogonal to each other, that is, they are mutually independent.

Species distribution models are not explicitly spatial (Franklin, 2009); they suppose that the geographical occurrences of records are mutually independent. However, this violates a fundamental principle of the spatial geography establishing that spatially proximate objects are similar and proximate localities tend to have similar values due to the possibility they reciprocally influence each other, or both are influenced by the same pattern that generates geographical processes (Franklin, 2009). Disregard and not avoid spatial autocorrelation has consequences, for example: *a)* it can increase the probability of incurring in type I errors or incorrectly rejecting the null hypothesis of no effect, *b)* variable selection may be predisposed toward more strongly auto-correlated predictors (Lennon, 2002), *c)* coarse scale predictors may be better selected against more locally influencing predictors, and *d)* model selection based on the Akaike information criterion will tend to model with larger number of predictive variables due to the committed residual variance structure. In summary, if spatial autocorrelation is present and ignored or not resolved, one may be incurring in a biased selection of variables or model-coefficients.

Among the statistical procedures proposed to solve or to reduce autocorrelation, principal component analysis (PCA), ridge-regression, and latent-root regression have been mentioned (Mason and Gunst, 1985; Afifi et al., 2012). The advantages of PCA compared with the other 2 procedures is the availability of an exact theory on estimate distributions, that is, the term or the error of the regression and the estimates are normally distributed (Gunst and Mason, 1977) and the principal components (PCs) are useful exploratory tools to detect and quantify mutual relationships among variables (Afifi et al., 2012).

The reduction of dimensionality is among the many applications of the PCA, that is, the reduction to a number of predictive variables that retain a high proportion of the original information (Tabachnick and Fidell, 2007). The PCs obtained are placed hierarchically according to their variance size; consequently, the first PC explains the maximum variance recorded in the predictive variables, the second PC explains the maximum of the residual variance and so forth, until the last PC which explains the remainder variance (Tabachnick and Fidell, 2007). Since the first PCs are those that retain the highest proportion