Una evaluación del sesgo de muestreo sobre el análisis ROC de modelos de nicho

Sergio Nicasio[1],[2]
Jean-François Mas[1]
Gabriela Hernández[1]

[1] Centro de Investigaciones en Geografía Ambiental
Universidad Nacional Autónoma de México
Campus Morelia 58190 – Morelia – Mich, México
jfmas@ciga.unam.mx

[2] Instituto de Investigaciones en Ecosistema y Sustentabilidad
Universidad Nacional Autónoma de México
Campus Morelia 58190 – Morelia – Mich, México
snicasio@cieco.unam.mx

Abstract. During the last decades, ecological niche modeling has become a very popular methodology in the exploration and analysis of biodiversity data. ROC analysis is widely used to assess the models and high performance is often reported in the literature. However, datasets derived from opportunistic observations often exhibit a strong geographic bias, mainly due to accessibility. This unequal coverage of a species distribution can strongly affect the quality of the model when important parts of the environmental space suitable to a specie are poorly represented in the survey dataset. This study aims at assessing the performance of ROC analysis in evaluating niche models. We elaborated independant niche models for *Romerolagus diazzi* using MaxEnt and data obtained during different decades separately. Each "decade based" model was trained using 75% of the data and assessed using the remaining 25%. ROC analysis based on the 25% of test data presented high scores for all the models. However, ACP analysis and the comparison between the species distribution derived from the models presented important differences. These results suggest that ROC analysis based on a subset of the data tend to be optimistically biased because the test set is not independant from the training set and presents often the same bias.

Keywords: niche modeling, ROC analysis, species distribution modelación de nicho, análisis ROC, distribución de especies

Introducción

En la última década la aplicación de los modelos de predicción de distribución de especies en ecología, evolución y biología de la conservación ha incrementado drásticamente (Lobo y Tognelli, 2011). Dichos modelos han sido indispensables para predecir la distribución potencial de las especies bajo diferentes escenarios climáticos (Anciaes y Peterson, 2006; Araújo et al., 2006), identificación de áreas vulnerables a especies invasoras (Anderson et al., 2006), procesos ecológicos como competencia entre especies (Anderson et al., 2002) y predecir sitios de riesgo a contraer enfermedades (López-Cárdenas et al., 2005; Reed et al., 2008), entre otras aplicaciones. Estas predicciones se obtienen a través de modelos lineales generalizados (Guisan et al., 2002), cálculo de distancias métricas (Larson y Olden, 2012), redes neuronales artificiales (Araújo et al., 2006; Broennimann et al., 2007), algoritmos genéticos (Anderson et al., 2002; López-Cárdenas et al., 2005; Anderson et al., 2006) y métodos de máxima entropía (Phillips et al., 2006; Reed et al., 2008).

Estos modelos presentan errores asociados a predecir probabilidades bajas donde existen registros observados de la especie (falso negativo o error de omisión), o probabilidades altas

en sitios poco adecuados para su presencia (falso positivo o error de comisión), afectando su verosimilitud. Los análisis ROC (receiver operating characteristic) permiten evaluar su poder predictivo a través del área bajo la curva (AUC) de la gráfica entre la sensibilidad (1 – tasa de error de omisión) y la 1 – especificidad (1 – tasa de error de comisión) del modelo. Un modelo con una AUC ≈ 1 tendrá un poder predictivo mayor respecto a uno con predicciones obtenidas de forma aleatoria (AUC ≤ 0.5). El cálculo del AUC requiere de registros independientes a aquellos empleados para elaborar el mapa de probabilidad del modelo. Una modificación de este método permite evaluar la validez estadística del modelo al modificar el tamaño del área bajo la curva dentro de una curva parcial (pROC).

Sin embargo, los resultados de estas evaluaciones deben considerarse cuidadosamente, debido a que los datos de presencia de las especies presentan diversas limitantes. Entre ellas están los errores de identificación taxonómica, la falta de inventarios completos, la sobre- o sub-representación de registros por especie y la obtención de datos a través de muestreos no sistematizados u oportunistas (Hijmans et al., 2000; Graham et al., 2004). Los sesgos de muestreo hacia áreas de fácil acceso o con un esfuerzo de muestreo mayor (cerca de caminos, en zonas de elevación alta o con baja pendiente), tienden a generar errores en la distribución predicha al sobrerepresentar presencia de una especie en esas áreas (Wisz et al., 2008). Es común encontrar modelos reportados con valores de AUC > 0.8 cuyos registros presentan estas limitantes.

El objectivo de este estudio es averiguar si las evaluaciones ROC, basadas en un subconjunto de los registros disponibles, sobreestiman el desempeño del modelo debido a que presentan los mismos sesgos que aquellos utilizados en la modelación.

1. Materiales y Métodos

Se obtuvieron registros de *Romerolagus diazzi* a través de la base de datos de la *Global Biodiversity Information Facility* (http://www.gbif.org/). Se eliminaron aquellos registros duplicados y que se encontraran fuera del área de estudio. Posteriormente, los registros depurados se dividieron en cuatro periodos de tiempo (1950-1970, 1970-1990, 1990-2010 y posterior a 2010). Posteriormente, se elaboraron los modelos de nicho para cada periodo de tiempo a través de MaxEnt con los criterios arriba mencionados, así como un análisis de componentes principales para evaluar la amplitud bioclimática de las variables más relevantes para cada periodo.

Los mapas de probabilidad de presencia de cada década se realizaron con el algoritmo de máxima entropía del programa MaxEnt y el paquete dismo, utilizando los parámetros que vienen por omisión en el programa, así como el 75% de los registros para el entrenamiento de los modelos y el 25% para su validación. Para reducir los efectos causados por la selección de puntos sobre los valores de probabilidad, se generaron cuatro réplicas a través del método de *Crossvalidate*. Los modelos obtenidos se validaron con el valor de AUC (*area under the curve*) del análisis ROC, el cuál indica si un modelo es mediana ($0.7 <$ AUC ≤ 0.9) o altamente predictivo (AUC > 9).

Para controlar los efectos que el uso de pseudoausencias y la ponderación homogénea de los errores de omisión y comisión tienen sobre los valores de AUC, estos modelos se evaluaron con el método de ROC parcial (pROC), utilizando el paquete *ENMGadgets*.

2. Resultados y discusión

La tabla 1 presenta los valores de AUC obtenidos para los modelos correspondientes a cada temporada. Se calculó el AUC utilizando los 25% de registros apartados para la validación (AUC y pAUC ”internas”) y utilizando todos los registros disponibles (de todas las décadas,

Table 1: AUC y AUC parcial de cada modelo basadas en los datos de validación de la década y en el conjunto de los datos disponibles.

Temporada	AUC ”interna”	AUC ”completa”	pAUC ”interna”	pAUC ”completa”
1970-1990	0.923	0.889	0.821	0.759
1990-2000	0.939	0.953	0.872	0.902
2000-2010	0.941	0.929	1.000	0.847
2010-2016	0.958	0.942	0.895	0.906

Table 2: Coincidencia entre los mapas de distribución de los diferentes modelos.

Temporada	1970	1990	2000	2010
1970	1			
1990	0.362	1		
2000	0.521	0.369	1	
2010	0.306	0.468	0.373	1

AUC y pAUC ”completas”). Se puede observar que todos los modelos presentan valores muy altos, entre 0.85 y 0.95, por lo cual se podría pensar que son una buena representación de la distribución real de la especie. Sin embargo, el análisis de componentes principales revela que las variables consideradas más importantes en la distribución de la especie difieren dependiendo de los datos utilizados (Figura 1).

De forma similar, las áreas de distribución obtenida de la umbralización de los mapas de probabilidad, aunque presentando un patrón general similar, tienen importantes diferencias (Figura 2). La mayor coincidencia ocurre entre los mapas de las décadas de 1970 y 2010 y solo alcanza 52%. La mayoría de los mapas tienen una coincidencia inferior a 40% (Tabla 2).

Estas contradicciones sugieren que los modelos elaborados para cada temporada tienden a ser evaluados de forma optimista porque esta evaluación se basa en datos, no utilizados directamente en el entrenamiento, pero no totalmente independientes de los datos utilizados para entrenar el modelo y presentando los mismos sesgos geográficos.

3. Conclusiones

La evaluación de los modelos de nicho es una tarea difícil porque los datos utilizados son a menudo escasos y presentan limitantes incluyendo errores de identificación taxonómica, errores de geolocalización y sesgos geográficos relacionados con muestreos no sistematizados u oportunistas (Hijmans et al., 2000; Graham et al., 2004). Los sesgos de muestreo hacia ciertas áreas, como las de fácil acceso, pueden generar errores en la distribución predicha al sobrerepresentar presencia de una especie en esas áreas (Wisz et al., 2008). Sin embargo, este efecto depende de la relación que existe entre las características ambientales de la distribución real de la especie y su representación en las áreas muestreadas. En este estudio, mostramos que modelos de la misma especie basados en datos diferentes tienen distribuciones geográficas muy constrastantes pero presentan todos altos valores de AUC. Por lo tanto, el análisis ROC no permitió una evaluación robusta de estos modelos.

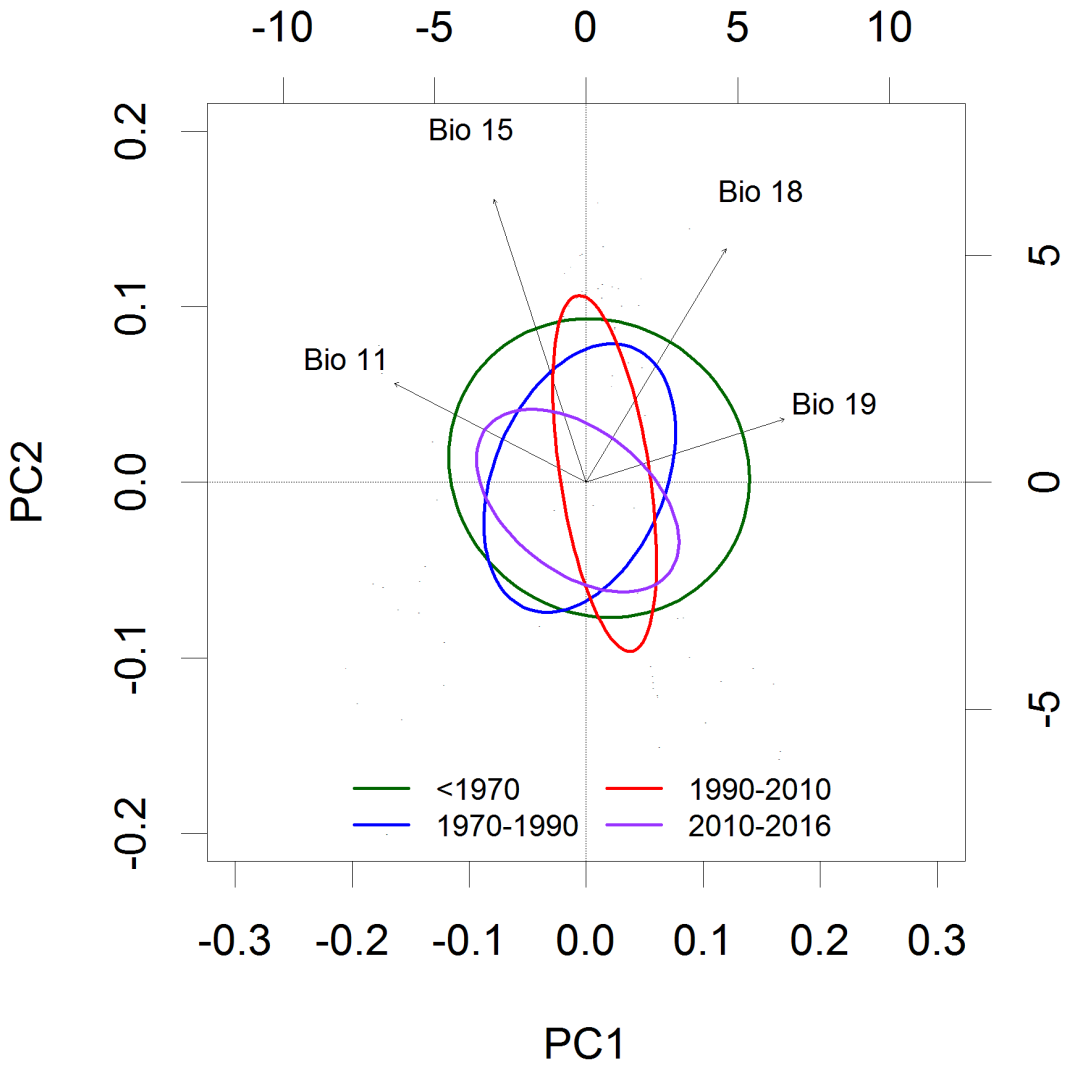

Figure 1: Análisis en componentes principales de los datos de cada temporada.

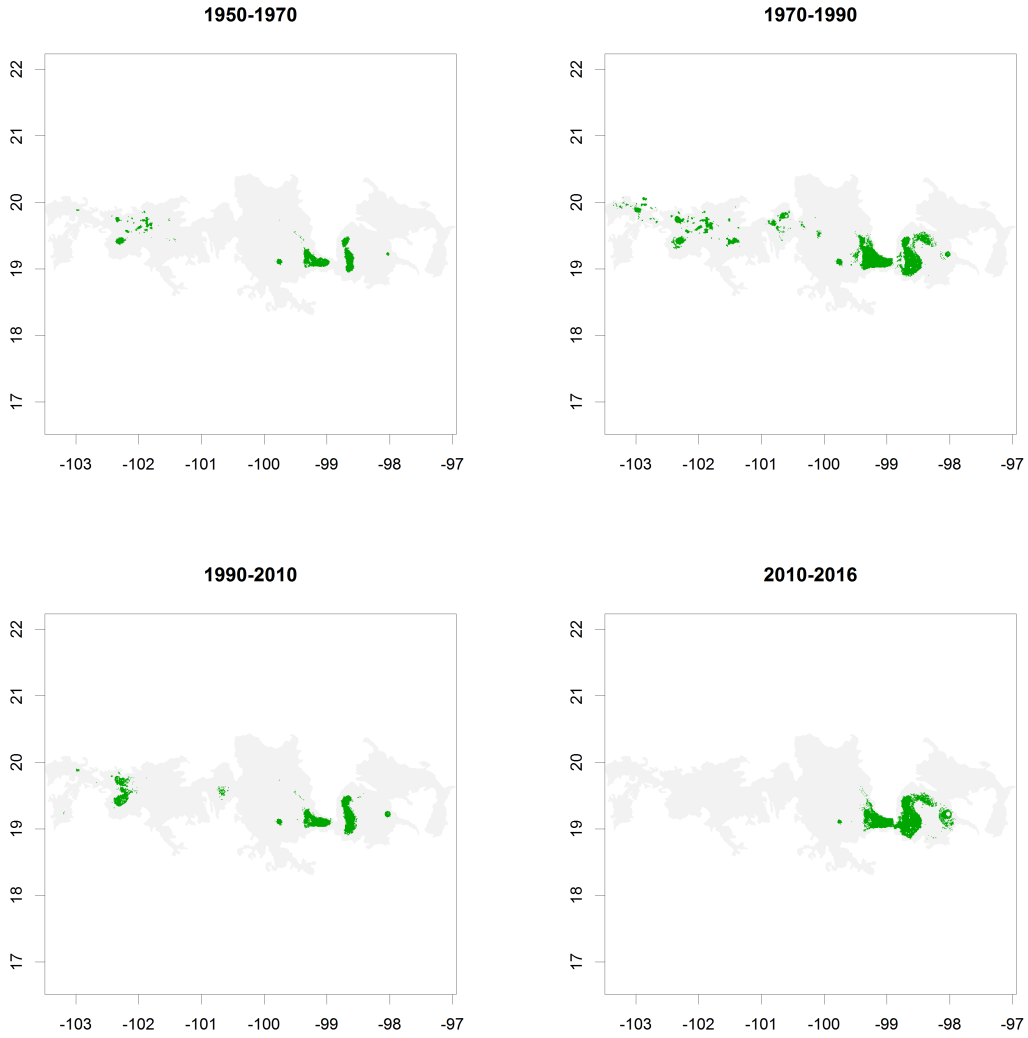

Figure 2: Mapas de distribución potencial de *Romerolagus diazzi* basados en los diferentes modelos por temporada.

Agradecimientos

Este trabajo se realizó con el apoyo del proyecto SEP-CONACYT 178816 *¿Puede la modelación espacial ayudarnos a entender los procesos de cambio de cobertura/uso del suelo y de degradación ambiental?*

Referencias

Anciães, M.; Peterson, A.T. Climate change effects on neotropical manakin diversity based on ecological niche modeling. **The Condor**, v. 108, p. 778–791, 2006.

Araújo, M.B.; Thuiller, W.; Pearson R.G. Climate warming and the decline of amphibians and reptiles in Europe. **J Biogeogr**, v. 33, p. 1712–1728, 2006.

Graham, C.H.; Ferrier, S.; Huettman, F., et al., New developments in museum-based informatics and applications in biodiversity analysis. **Trends Ecol Evol**, v. 19, p. 497–503, 2004.

Hijmans, R.J.; Garrett, K.A.; Huamán, Z. et al., Assessing the Geographic Representativeness of Genebank Collections: the Case of Bolivian Wild Potatoes. **Conserv Biol**, v. 14, p. 1755–1765, 2000.

Lobo, J.M.; Jiménez-Valverde, A;, Real, R. AUC: a misleading measure of the performance of predictive distribution models. **Glob Ecol Biogeogr**, v. 17, p. 145–151, 2008.

López-Cárdenas, J., Gonzalez Bravo, F.E.; Salazar Schettino, P.M. et al. Fine-scale predictions of distributions of Chagas disease vectors in the state of Guanajuato, Mexico. **J Med Entomol**, v. 42, p. 1068–1081, 2005.

Peterson, A.T. **Ecological niches and geographic distributions**. Princeton University Press, 2011, 314 p.

Peterson, A.T.; Papeş, M.; Soberón, J. Rethinking receiver operating characteristic analysis applications in ecological niche modeling. **Ecol Model**, v. 213, p. 63–72, 2008.

Reed, K.D.; Meece, J.K.; Archer, J.R.; Peterson, A.T. Ecologic Niche Modeling of Blastomyces dermatitidis in Wisconsin. **PLOS ONE**, v. 3:e2034, 2008.

Wisz, M.S.; Hijmans, R.J.; Li, J. et al. Effects of sample size on the performance of species distribution models. **Divers Distrib**, v. 14, p. 763–773, 2008.