

IDENTIFICATION OF AREAS OF ENDEMISM FROM SPECIES DISTRIBUTION MODELS: THRESHOLD SELECTION AND NEARCTIC MAMMALS

Tania Escalante^{1*}, Gerardo Rodríguez-Tapia², Miguel Linaje³,
Patricia Illoldi-Rangel⁴ and Rafael González-López¹

¹Museo de Zoología "Alfonso L. Herrera", Depto. de Biología Evolutiva, Facultad de Ciencias, Universidad Nacional Autónoma de México. Apdo. Postal 70-399, C.P. 04510, México, D.F. ²Unidad de Geomática, Instituto de Ecología, Universidad Nacional Autónoma de México. ³Lab. de Sistemas de Información Geográfica, Depto. de Zoología, Instituto de Biología, Universidad Nacional Autónoma de México. ⁴Biodiversity and Biocultural Conservation Laboratory, Section of Integrative Biology, University of Texas at Austin. E-mail: ^{1*}tee@ibunam2.ibiologia.unam.mx, jrglezlpz@gmail.com, ²gerardo@ecologia.unam.mx, ³miglinaje@gmail.com, ⁴patz30@gmail.com

ABSTRACT

We evaluated the relevance of threshold selection in species distribution models on the delimitation of areas of endemism, using as case study the North American mammals. We modeled 40 species of endemic mammals of the Nearctic region with Maxent, and transformed these models to binary maps using four different thresholds: minimum training presence, tenth percentile training presence, equal training sensitivity and specificity, and 0.5 logistic probability. We analyzed the binary maps with the optimality method in order to identify areas of endemism and compare our results regarding previous analyses. The majority of the species tend to have very low values for the minimum training presence, whereas most of the species have a value of the tenth percentile training presence around 0.5, and the equal training sensitivity and specificity was around 0.3. Only with the tenth percentile threshold we recovered three out of the four patterns of endemism identified in North America, and detected more endemic species. The best identification of areas of endemism was obtained using the tenth percentile training presence threshold, which seems to recover better the distributional area of the mammals analyzed.

Key Words: Analysis of endemism, Mammalia, Maxent, Nearctic region, optimality.

RESUMEN

Evaluamos la relevancia de la selección del umbral en los modelos de distribución de especies en la delimitación de las áreas de endemismo, usando como un caso de estudio a los mamíferos de América del Norte. Modelamos 40 especies de mamíferos endémicos de la región Neártica con Maxent, y transformamos esos modelos a mapas binarios usando cuatro umbrales diferentes: presencia mínima de entrenamiento, percentil diez de la presencia de entrenamiento, igual sensibilidad y especificidad de entrenamiento, y probabilidad logística de 0.5. Los mapas binarios los analizamos con el método de optimación con el objeto de identificar áreas de endemismo y comparar nuestros resultados con estudios previos. La mayoría de las especies mostraron tendencias hacia valores muy bajos de la presencia mínima de entrenamiento, mientras que la mayoría tuvo un valor del percentil diez de la presencia de entrenamiento alrededor de 0.5, y de igual sensibilidad y especificidad de entrenamiento alrededor de 0.3. Únicamente con el percentil diez de la presencia de entrenamiento se recuperaron tres de los cuatro patrones de endemismo identificados para América del Norte y se detectaron más especies endémicas. La identificación de áreas de endemismo más eficiente se obtuvo usando el umbral del percentil diez de la presencia de entrenamiento, el cual parece recuperar mejor las áreas de distribución de los mamíferos analizados.

Palabras Clave: Análisis de endemismo, Mammalia, Maxent, región Neártica, optimación.

INTRODUCTION

Species distribution models (also named ecological niche models) are commonly used in biogeography. In particular, although they are more suited for the identification of ecological biogeographical patterns, they also have important applications in the identification of historical biogeographical patterns, namely, generalized tracks¹ and areas of endemism²⁻⁶ where models have been used to improve their delimitation.

There are many modeling techniques (GLM, GAM, GARP, ENFA, Maxent, etc.), which can be used depending on the available records (data) for each species, environmental data and the required accuracy of the models. Some comparisons of the different modeling techniques have been performed⁷⁻⁹ and although there are no general conclusions, Maxent¹⁰⁻¹² seems to perform better than others. Maxent generates probability maps of species presence in three output formats: raw, cumulative and logistic (see Maxent tutorial, <http://www.cs.princeton.edu/~schapire/maxent/>), being the last two the most used (in scales of 0-100 and 0-1, respectively).

As in conservation and environmental management practices¹³, in biogeography sometimes it is necessary to transform probabilistic data to presence/absence data (binary maps, i.e. 1 - 0). For this to be feasible, a probability threshold has to be established to the minimum level at which the distributions should be left out. As there are many possible uses for distribution models, some methods have been proposed in order to select the best threshold in Maxent to obtain a binary map for species (see Table I). They include the minimum (or lowest) training presence, threshold of a particular percentage (10, 50, 80%), sensitivity at 95%, some percentile training presence (10, 20), equal training sensitivity and specificity, etc. (Pawar *et al.*¹⁴ for further details). However, there has been some comparisons and evaluations that might allow to select the best threshold for other modeling algorithms generally related with prevalence, sensitivity and specificity^{13,15-17}, and specifically for Maxent¹⁸⁻²⁰ (see Table I). So, there is not a consensus about which is the way to select the best threshold.

Areas of endemism are basic biogeographic units, their identification is the first step of an evolutionary biogeographic analysis and they are a pre-requisite of any cladistic biogeographic analysis²¹. An area of endemism is an area of non-random distributional congruence of two or more taxa²², and the basis of biogeographic regionalizations²³. The identification of areas of endemism depends totally on maps of distribution of species and their generalization to spatial units. The most used units of study are grid-cells, although it is possible to use other regular polygons or even polygons with irregular forms. The most popular methods (Parsimony Analysis of Endemism²¹ and Endemism analysis^{24,25}) employ data matrices of presence/absence of species in quadrats. Thus, the identification of areas of endemism can be affected by

the generalization of individual areas of distribution to the grid-cells. Some authors^{6,26} pointed that the use of species distribution models (or ecological niche models) can modify the identification of areas of endemism due to the overprediction involved in them; however, this has not been proved.

Escalante *et al.*²⁷ recently published a study of identification of Nearctic areas of endemism using mammals. They used areas of distribution drawn by traditional methodology (areas inferred by mammalogists specialists; maps available on <http://conabiweb.conabio.gob.mx/website/mamiferos/viewer.htm>²⁸), in order to analyze the main patterns of endemism corresponding to the Nearctic region. They obtained four areas in North America identified by 40 species: Nearctic, Western, Eastern and Northern patterns.

We evaluate herein the relevance of the selection of the threshold in Maxent using four different options (minimum training presence, tenth percentile training presence, equal training sensitivity and specificity and 0.5 logistic probability), and its impact on the delimitation of areas of endemism, using as study case the mammals of the Nearctic region.

MATERIAL AND METHODS

We compiled a database of 40 species of endemic mammals of North America (following Escalante *et al.*²⁷) corresponding to five orders (Table II). Those species gave score to some area of endemism in that publication, and shown sympatric patterns. Records were obtained from a database of mammals of Mexico (Mammex; Escalante *et al.*, unpublished data), and four on-line databases: GBIF (<http://www.gbif.org/>), MaNIS (<http://manisnet.org/>), CONABIO (Remib; <http://www.conabio.gob.mx/>), and UNIBIO (Instituto de Biología, UNAM; <http://unibio.ibiologia.unam.mx/>). A record is considered as a unique combination of the name of the species and georeferenced site (latitude-longitude) (Table II). Localities of each species were geographically validated in a Geographic Information System (GIS; ArcGis 9.3)²⁹, using specialized bibliography^{30,31} and two websites: North American Mammals (<http://www.mnh.si.edu/mna/>) and Infonatura (<http://www.natureserve.org/infonatura/>).

To construct the models in Maxent, 23 environmental data layers were used at a resolution of ~2 km (which is suitable for our study area): four topographic layers were obtained from Hydro1k (<http://edc.usgs.gov/products/elevation/gtopo30/hydro/america.html>) while 19 climatic data layers were derived from the WorldClim database (<http://www.worldclim.org/>: altitude, aspect, compound topographic index, slope, annual mean temperature, mean diurnal range, isothermality, temperature seasonality, maximum temperature of warmest month, minimum temperature of coldest month, temperature annual range, mean temperature of wettest quarter, mean temperature of driest quarter, mean temperature of warmest quarter, mean temperature

Reference	Criteria	Taxa and data
Papes & Gaubert (2007) ³³	(Maxent 0 to 100) All probability values >0.	Mammals. Museum collections, databases and literature.
Pearson <i>et al.</i> (2007) ¹⁸	(Maxent 0 to 100) Lowest presence threshold and threshold 10.	Geckos. Museum collections.
Loiselle <i>et al.</i> (2008) ³⁴	(Maxent 0 to 100) Threshold of 1 in all Maxent predictions of species distributions. When the prediction value was equal to or above 1, predicted the presence of the species. A value of 1 was sufficient to capture all of the presence training points within the predicted distribution.	Plant species. Herbarium collections.
Waltari & Guralnick (2009) ³⁵	(Maxent 0 to 100) Modified lowest-presence threshold (95% of all occurrences in the training dataset falling into suitable habitat, representing a less stringent model); and threshold 50 (representing a more stringent threshold).	Mammals. Museum collections.
Costa <i>et al.</i> (2009) ³⁶	Lowest presence threshold and Parameter <i>E</i> (measure of the amount of error associated with the presence localities dataset) at 5%.	Reptiles. Museum collections, literature and fieldwork.
Brito <i>et al.</i> (2009) ³⁷	The tenth percentile training presence thresholds were chosen because 'true' absence data was not available. Models were reclassified with "Reclassify" function of ArcMap.	Canids. Observations, bibliography and museum collections. "Nearest Neighbour Index" of ArcMap GIS assessed the degree of clustering of the data.
Newbold <i>et al.</i> (2009) ³⁸	Threshold that resulted in a sensitivity of 95%.	Butterflies and mammals. Museum specimens and literature.
Ramírez-Barahona <i>et al.</i> (2009) ¹	(Maxent 0 to 100). Threshold of 80: pixels with a maximum entropy value of less than 80 were eliminated.	Plant species (ferns and lycopods). Herbarium collections.
Colacicco-Mayhugh, Masuoka & Grieco (2010) ³⁹	Minimum training presence.	Diptera. Literature and collection records.
Donegan & Avendaño (2010) ⁴⁰	20th percentile training presence.	Birds. Field and collection records.
Giovanelli <i>et al.</i> (2010) ⁴¹	Minimum presence threshold, that equals the minimum model prediction value for any of the training occurrence point data.	Anura (Hylidae). Precise and uniform sampling (none of the occurrences should be an outlier in environmental space)
Torres & Jayat (2010) ⁴²	Maximum training sensitivity and specificity and average of values of all pixels with prediction.	Four species of mammals. Field and collection records.
Aranda & Lobo (2011) ¹⁹	21 decision thresholds were selected at intervals of 5 to 100, and minimum training presence.	Plant species. Database.

Table I. Some thresholds for Maxent to transform to binary maps, using different taxa and origin of data. For the criteria described in this table, sensitivity refers to the proportion of presences correctly predicted. Specificity is the proportion of absences correctly predicted. Both are indices, not criteria. Prevalence refers to the proportion of the study area covered by the species' distributional area¹³.

of coldest quarter, annual precipitation, precipitation of wettest month, precipitation of driest month, precipitation seasonality, precipitation of wettest quarter, precipitation of driest quarter, precipitation of warmest quarter and precipitation of coldest quarter.

For each species, 25% of the records were used to validate the model internally. The algorithm of Maxent uses a series of rules to calculate probabilities. For the present analysis, all rules were used, so the program selects the adequate one depending on the number of available data. The used rules are: (a) linear, which uses the variable by itself; (b) quadratic, which uses the square

Order/Species	Number of records		AUC		Threshold		
	(a)	(b)	(a)	(b)	(a)	(b)	(c)
Carnivora							
<i>Canis rufus</i>	23	7	0.998	0.960	0.312	0.467	0.312
<i>Martes americana</i>	336	111	0.973	0.953	0.020	0.419	0.397
Lagomorpha							
<i>Brachylagus idahoensis</i>	66	21	0.992	0.988	0.029	0.374	0.208
<i>Lepus americanus</i>	199	66	0.957	0.931	0.036	0.306	0.271
<i>Ochotona princeps</i>	151	50	0.996	0.988	0.019	0.525	0.274
<i>Sylvilagus aquaticus</i>	128	42	0.997	0.992	0.033	0.456	0.198
<i>Sylvilagus nuttallii</i>	51	17	0.992	0.992	0.055	0.360	0.193
Soricomorpha							
<i>Blarina carolinensis</i>	64	21	0.986	0.957	0.007	0.382	0.199
<i>Sorex cinereus</i>	771	256	0.943	0.915	0.007	0.383	0.428
<i>Sorex longirostris</i>	16	5	0.990	0.965	0.093	0.209	0.093
<i>Sorex merriami</i>	40	13	0.994	0.993	0.031	0.404	0.105
<i>Sorex palustris</i>	83	27	0.973	0.912	0.101	0.287	0.276
Chiroptera							
<i>Crynorhinus rafinesquii</i>	9	3	0.990	0.997	0.247	0.247	0.247
<i>Lasiurus seminolus</i>	98	32	0.998	0.995	0.255	0.546	0.300
<i>Myotis austroriparius</i>	59	19	0.991	0.994	0.039	0.391	0.233
<i>Myotis sodalis</i>	67	22	0.998	0.978	0.140	0.239	0.180
<i>Nycticeius humeralis</i>	234	78	0.986	0.980	0.129	0.439	0.345
Rodentia							
<i>Erethizon dorsata</i>	482	160	0.940	0.880	0.015	0.387	0.440
<i>Lemmys curtatus</i>	164	54	0.992	0.989	0.059	0.416	0.235
<i>Lemmus sibiricus</i>	42	13	0.972	0.867	0.173	0.332	0.325
<i>Marmota flaviventris</i>	522	173	0.987	0.983	0.003	0.469	0.388
<i>Microtus montanus</i>	729	242	0.986	0.985	0.014	0.479	0.408
<i>Microtus pennsylvanicus</i>	1322	440	0.917	0.900	0.009	0.408	0.486
<i>Microtus pinetorum</i>	277	92	0.987	0.978	0.040	0.459	0.389
<i>Microtus richardsoni</i>	129	43	0.995	0.988	0.009	0.428	0.183
<i>Myodes rutilus</i>	27	9	0.969	0.945	0.053	0.309	0.302
<i>Ochrotomys nuttalli</i>	176	58	0.993	0.984	0.048	0.514	0.363
<i>Oryzomys palustris</i>	225	75	0.994	0.990	0.062	0.486	0.342
<i>Perognathus parvus</i>	605	201	0.993	0.990	0.048	0.523	0.345
<i>Peromyscus gossypinus</i>	403	134	0.992	0.992	0.029	0.490	0.351
<i>Reithrodontomys humulis</i>	66	21	0.989	0.989	0.010	0.359	0.279
<i>Spermophilus columbianus</i>	165	55	0.994	0.991	0.061	0.538	0.278
<i>Spermophilus elegans</i>	44	14	0.991	0.984	0.020	0.303	0.085
<i>Spermophilus lateralis</i>	306	101	0.995	0.992	0.096	0.482	0.327
<i>Spermophilus parryii</i>	244	81	0.969	0.954	0.048	0.381	0.355
<i>Tamias amoenus</i>	980	326	0.988	0.988	0.015	0.496	0.377
<i>Tamias ruficaudus</i>	107	35	0.998	0.996	0.193	0.600	0.355
<i>Tamiasciurus hudsonicus</i>	2019	627	0.936	0.930	0.002	0.410	0.482
<i>Thomomys talpoides</i>	1161	386	0.978	0.976	0.026	0.483	0.447
<i>Thomomys townsendii</i>	99	33	0.999	0.999	0.014	0.664	0.329

Table II. Data of the models for endemic species. Number of records: (a) used in the training of models and (b) in the test of models; the AUC for: (a) training and (b) testing; and the value of the threshold for logistic models: (a) minimum training presence, and (b) the tenth percentile training presence, and (c) equal training sensitivity and specificity.

of the variable; (c) product, which uses the product of two variables; (d) threshold, which uses a binary transformation (0, 1) of a continuous variable using a threshold; and (e) hinge, which is like the lineal rule, but remains constant under the threshold. The algorithm determines which rule to use like follows: lineal if there are < 10 points; lineal + cuadratic if there are 10-14 points; lineal + cuadratic + hinge if there are 15-79 points; and all if there are > 80 points (<http://www.cs.princeton.edu/~schapire/maxent/tutorial/tutorial.doc>). The logistic value output was selected because is the easiest to conceptualize since it gives an estimate between 0 and 1 of probability of presence (see <http://www.cs.princeton.edu/~schapire/maxent/tutorial/tutorial.doc> for further details).

Model success was judged using two criteria: $AUC > 0.7$, and $p < 0.05$ for at least one binomial test¹⁴, and both obtained from the program. AUC, or area under the curve, is an index used to evaluate models because it provides a single measure of overall accuracy that is not dependent upon a particular threshold⁴³. The value of the AUC ranges between 0 and 1.0. Values of 0.5 implies that the scores for two groups (random and model) do not differ, while a score of 1.0 indicates no overlap in the distributions, and the model is reliable. A value of 0.8 for the AUC means that for 80% of the time a random selection from the positive group will have a score greater than a random selection from the negative class. It is important to note that AUC values tend to be higher for species with narrow ranges, relative to the study area described by the environmental data. This does not necessarily mean that the models are better; instead this behavior is an artifact of the AUC statistic⁴³.

Models were generated in ascii format, and exported directly to the GIS. We selected four of the most common used thresholds for Maxent models in logistic format: the minimum training presence, the tenth percentile training presence, the equal training sensitivity and specificity (obtained from the output table of Maxent), and a logistic probability of 0.5. All pixels with a value under those thresholds were assigned a value of zero (0), which would represent absence of the species.

To analyze the influence of the four thresholds on the delimitation of areas of endemism, the 40 endemic species were analyzed, in order to prove if we identify the patterns previously discovered²⁷. We overlapped and intersected the binary maps obtained for each species, using each one of the four thresholds (minimum training presence, tenth percentile training presence, equal training sensitivity and specificity and logistic probability of 0.5) to a 4° latitude-longitude grid. Then, we built four matrices of presence/absence (one for each threshold), where the predicted presence of a species was coded as "1" and its absence was coded as "0". We performed four analysis of endemicity with the optimality method^{24,25}, one for each threshold. The optimality method calculates a score of endemicity for a taxon to a given area (grid), so, the endemicity for an area will be the sum of the scores

of two or more taxa inhabiting it. From among different possible areas, those with the highest scores of endemicity are preferred.

The four analyses of endemicity were developed in NDM/VNDM v. 2.5⁴⁴ (available at www.zmuc.dk/public/phylogeny), where each matrix was analyzed iteratively changing the random seed until the number of areas of endemism remained stable. We used the same parameters used by Escalante *et al.*²⁷: heuristic search saving sets of areas with two or more endemic species, save sets with score above 2, and optimal sets were chosen when having above 50% of different endemic species to the highest score. When we obtained two or more areas of endemism, consensus areas were calculated using 30% of similarity in species against any of the other areas in the consensus. We obtained the number of endemic taxa of each matrix and their consensus areas of endemism. All areas of endemism were analyzed regarding their scores, patterns represented and number of endemic species, in order to compare them with the analysis of Escalante *et al.*²⁷ and to evaluate the performance of the four thresholds.

RESULTS

We obtained 40 models from Maxent (one for each species). The average value for the AUC for training was 0.98 and 0.96 for testing (see Table II). The values for the minimum training presence, the tenth percentile training presence and the equal training sensitivity and specificity thresholds for each species are shown in Table II. The range for the minimum training presence was 0.002 - 0.312, for the tenth percentile presence was 0.209 - 0.664, and for the equal training sensitivity and specificity was 0.085-0.486, with averages of 0.065, 0.412, and 0.303, respectively. Most of the species tend to have very low values for the minimum training presence, whereas most of species have a value of the tenth percentile training presence around of 0.5, and the equal training sensitivity and specificity less than 0.5. An example of the differences between the binary maps resulting from the application of four thresholds is shown in Figures 1 and 2.

The results of the analyses of endemicity are shown in Tables III and IV. In the analysis using the minimum training presence threshold, we could recover only one pattern of endemism (Fig. 3): the Western pattern of Escalante *et al.*²⁷ With the tenth percentile threshold we recovered three patterns (Fig. 4): Nearctic, Western and Eastern; with the 0.5 value of probability as a threshold, we recovered two patterns (Fig. 5): Western and Eastern; and the same with the equal training sensitivity and specificity, two patterns were identified: Western and Eastern (Fig. 6). Moreover, the threshold where we obtained more endemic species was the tenth percentile, followed by the 0.5, the equal training sensitivity and specificity and the minimum training presence (Table IV). Only one pattern (the Northern pattern) of Escalante *et al.*²⁷ could not be recovered with any of the thresholds.

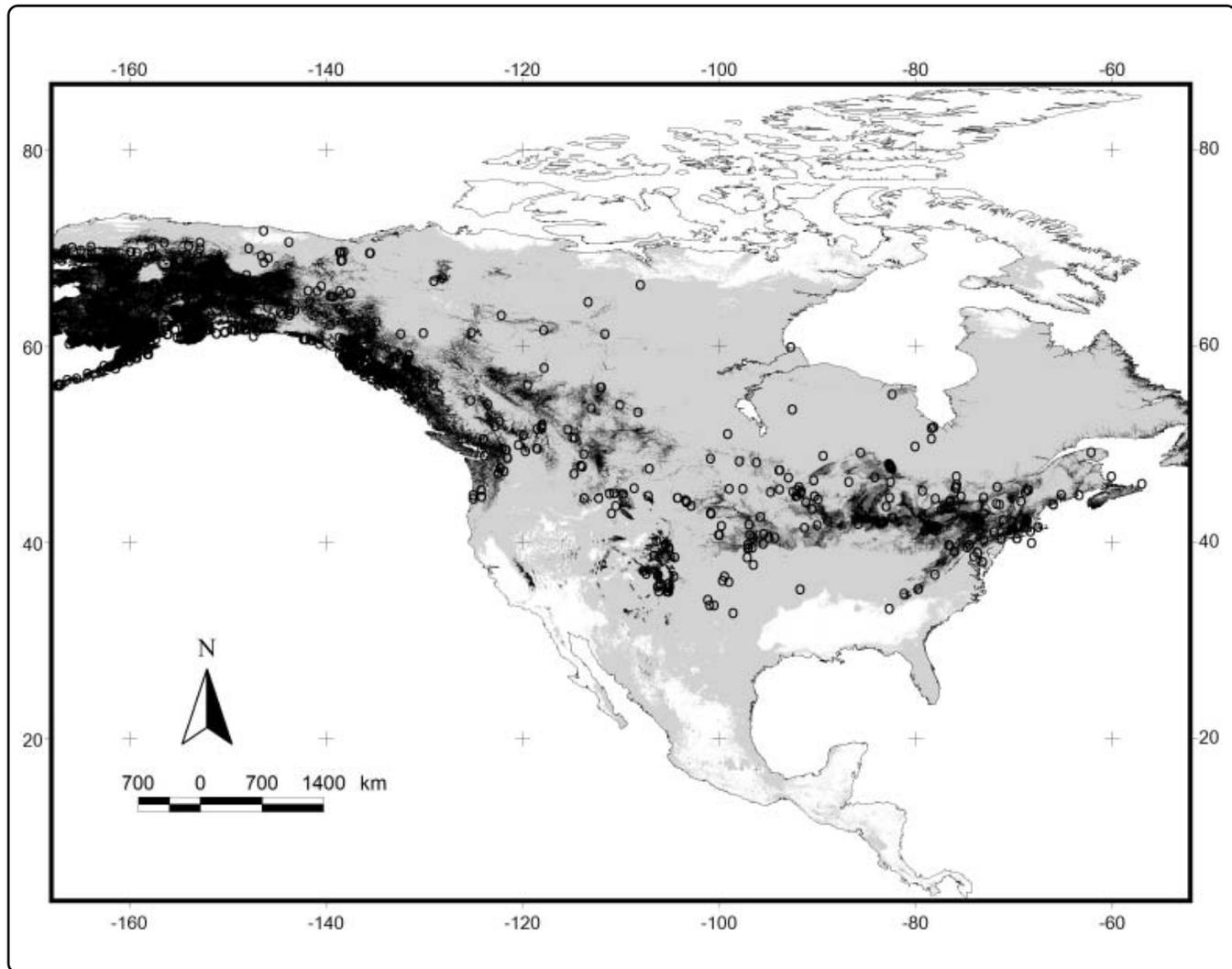


Figure 1. Map of potential distribution of *Sorex cinereus* in North America with four different thresholds: black, the probability of 0.5; dark gray, the tenth percentile training presence (0.383); medium gray, the equal training sensitivity and specificity (0.428); and light gray, the minimum training presence (0.007). Circles: data points.

Threshold	Number of areas of endemism	Number of consensus areas	Number and name of general patterns represented	Number of endemic species	Range of scores of consensus areas
Minimum training presence	1	1	1 – Western pattern	3	2.6096
0.5	4	4	2 – Western and Eastern patterns	19	2.0811-7.0542
Equal training sensitivity and specificity	3	2	2 – Western and Eastern patterns	14	3.5820-5.5790
Tenth percentile training presence	4		3 – Western, Eastern and Nearctic patterns	22	2.3135-7.3247

Table III. Areas of endemism and consensus areas for each threshold.

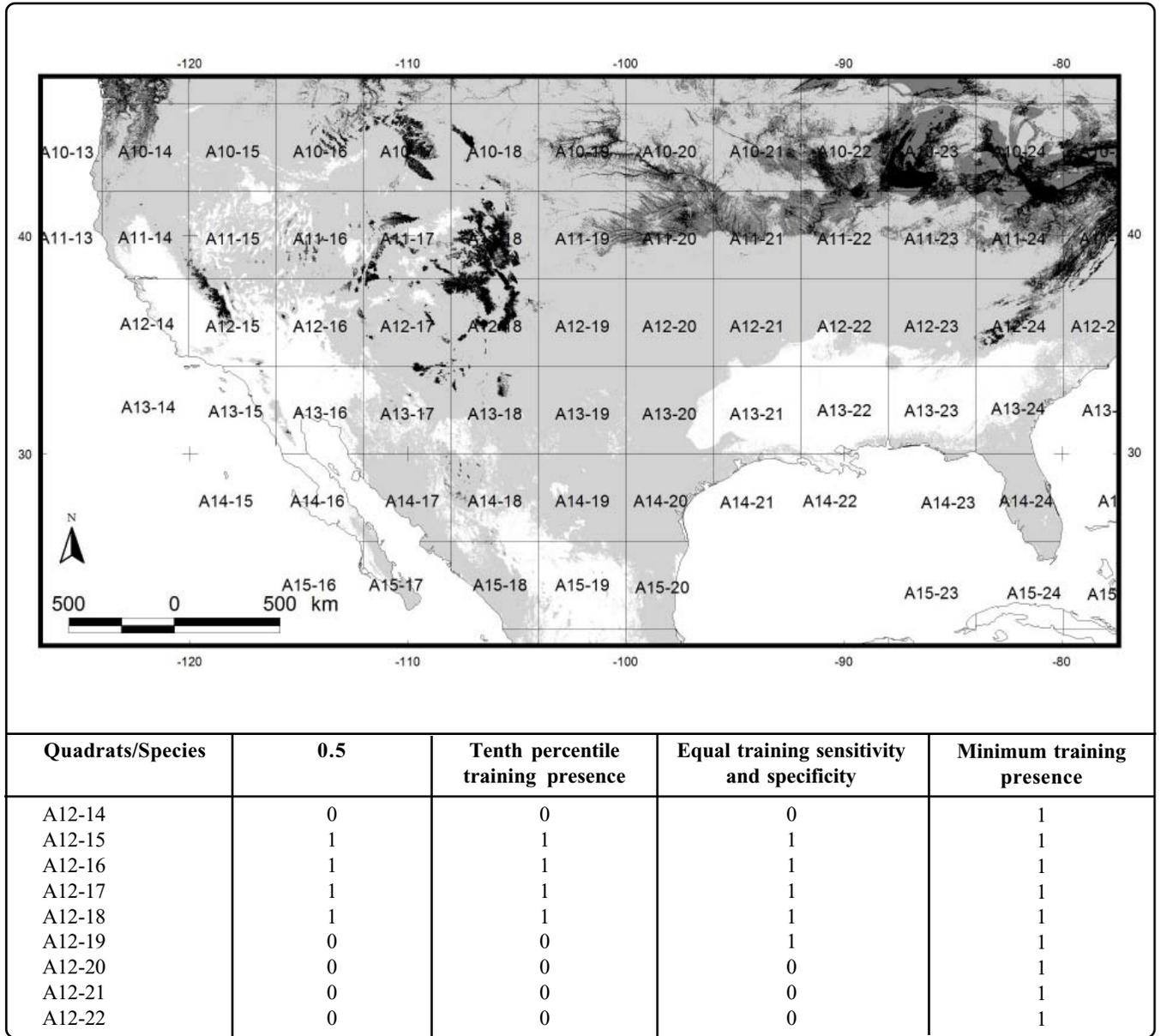


Figure 2. Detail of a generalization of the four potential distributional areas of *Sorex cinereus* to a 4° grid on the Mexico-U.S.A. border. The presence predicted by each map in a quadrat is coded with "1", and the absence with "0". The label of each 4° quadrat is showed as A#-#. Black: the probability of 0.5; dark gray: the tenth percentile training presence (0.383); medium gray: the equal training sensitivity and specificity (0.428); and light gray: the minimum training presence (0.007).

Species	Order	Minimum training presence	0.5	Tenth percentile training presence	Equal training sensitivity and specificity
Nearctic region					
<i>Erethizon dorsatum</i>	Rodentia				
<i>Lepus americanus</i>	Lagomorpha				
<i>Microtus pennsylvanicus</i>	Rodentia			X	
<i>Sorex cinereus</i>	Soricomorpha			X	
<i>Tamiasciurus hudsonicus</i>	Rodentia			X	
<i>Sorex palustris</i>	Soricomorpha				
<i>Martes americana</i>	Carnivora				
Western pattern					
<i>Brachylagus idahoensis</i>	Lagomorpha	X	X	X	X
<i>Lemmiscus curtatus</i>	Rodentia		X	X	
<i>Marmota flaviventris</i>	Rodentia		X	X	X
<i>Microtus montanus</i>	Rodentia		X	X	X
<i>M. richardsoni</i>	Rodentia		X	X	
<i>Ochotona princeps</i>	Lagomorpha			X	
<i>Perognathus parvus</i>	Rodentia	X	X	X	X
<i>Sorex merriami</i>	Soricomorpha		X		
<i>Spermophilus columbianus</i>	Rodentia		X	X	
<i>Spermophilus elegans</i>	Rodentia			X	
<i>Spermophilus lateralis</i>	Rodentia				X
<i>Sylvilagus nuttallii</i>	Lagomorpha				
<i>Tamias amoenus</i>	Rodentia		X	X	X
<i>Tamias ruficaudus</i>	Rodentia	X	X	X	X
<i>Thomomys talpoides</i>	Rodentia				
<i>Thomomys townsendii</i>	Rodentia				X
Eastern pattern					
<i>Blarina carolinensis</i>	Soricomorpha		X	X	
<i>Canis rufus</i>	Carnivora		X	X	X
<i>Corynorhinus rafinesquii</i>	Chiroptera				
<i>Lasiurus seminolus</i>	Chiroptera		X	X	
<i>Microtus pinetorum</i>	Rodentia				
<i>Myotis austroriparius</i>	Chiroptera			X	
<i>Myotis sodalis</i>	Chiroptera				
<i>Nycticeius humeralis</i>	Chiroptera				
<i>Ochrotomys nuttalli</i>	Rodentia		X	X	X
<i>Oryzomys palustris</i>	Rodentia		X	X	X
<i>Peromyscus gossypinus</i>	Rodentia		X	X	X
<i>Sorex longirostris</i>	Soricomorpha				
<i>Sylvilagus aquaticus</i>	Lagomorpha		X		X
<i>Reithrodontomys humulis</i>	Rodentia		X	X	X
Northern pattern					
<i>Clethrionomys rutilus</i>	Rodentia				
<i>Lemmus sibiricus</i>	Rodentia				
<i>Spermophilus parryii</i>	Rodentia				

Table IV. Results of the analyses of endemicity for 40 endemic species of the Nearctic region for three thresholds. X= species recovered in each analysis.

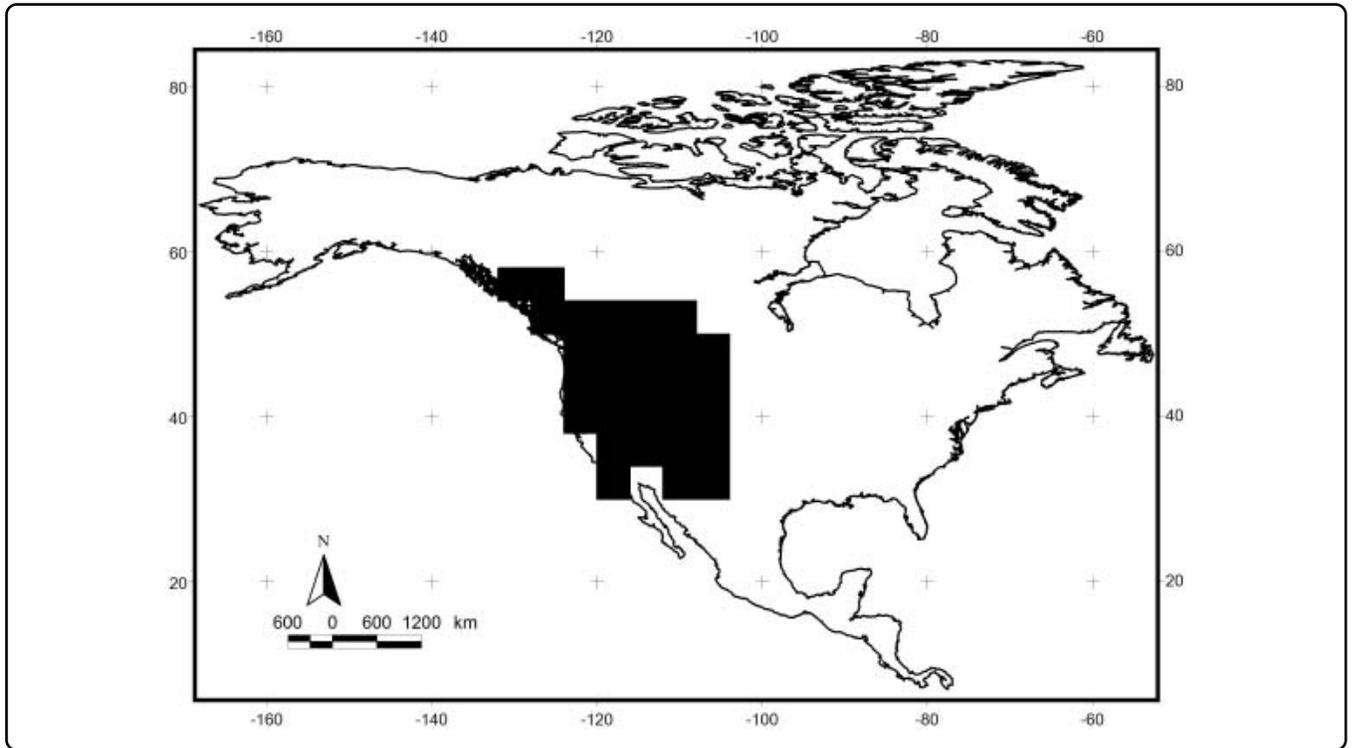


Figure 3. Area of endemism in North America obtained from the matrix with the minimum training presence threshold. Black quadrats: Western pattern.

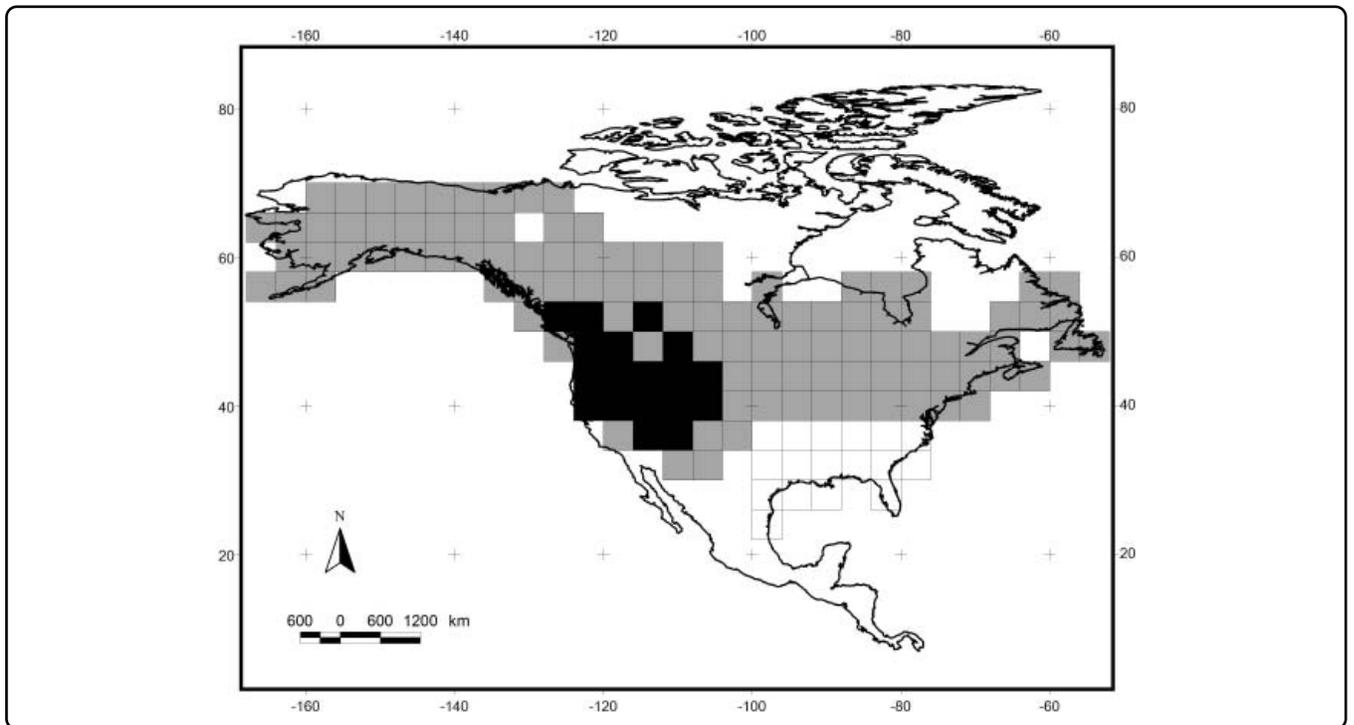


Figure 4. Three areas of endemism in North America obtained from the matrix with the tenth percentile training presence threshold. Black quadrats: Western pattern; gray quadrats: Nearctic pattern; white quadrats: Eastern pattern.

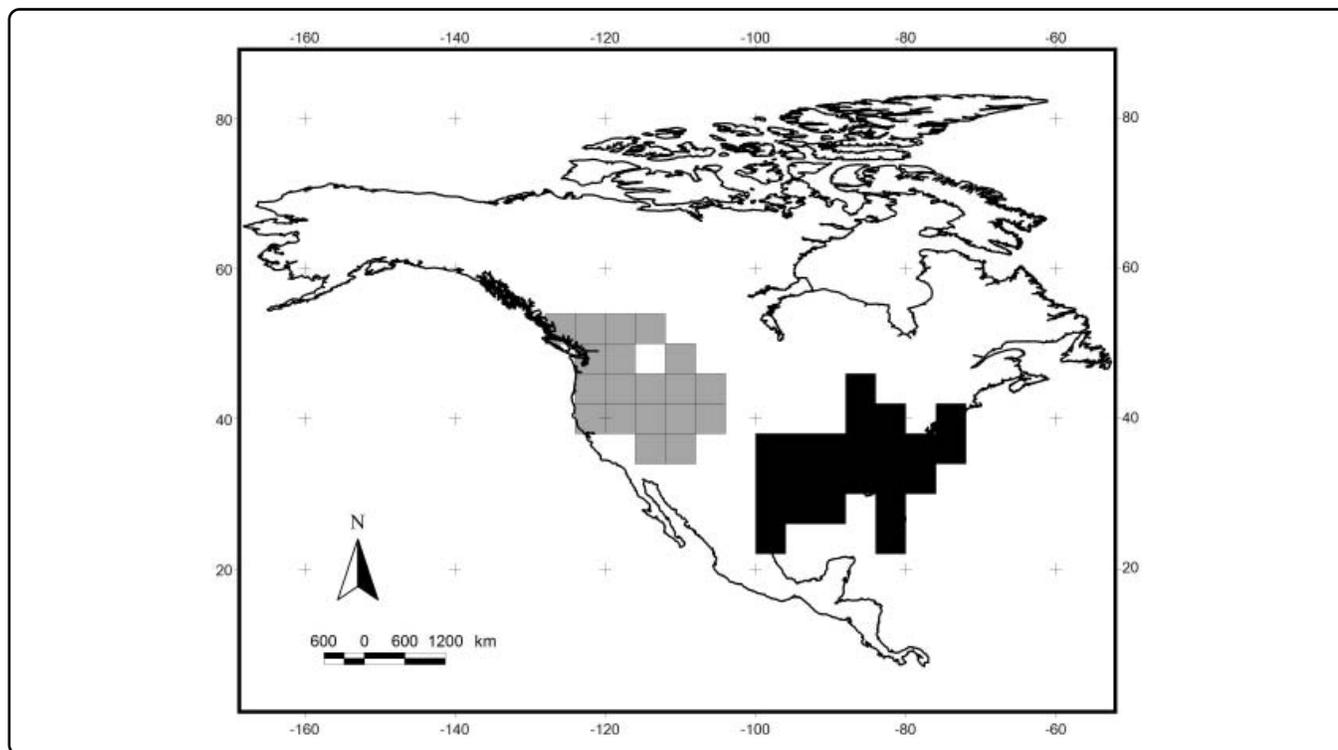


Figure 5. Two patterns of endemism in North America obtained from the matrix with the 0.5 threshold. Gray quadrats: Western pattern; white quadrats: Eastern pattern.

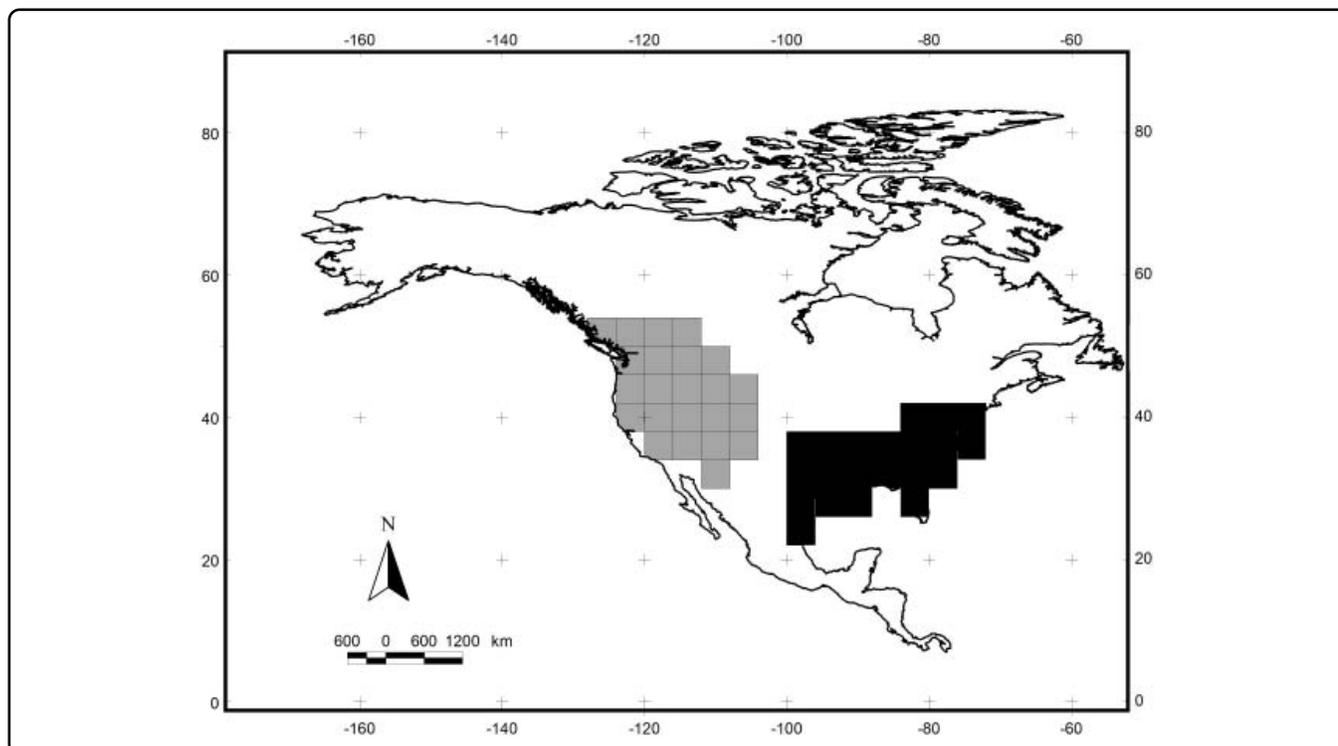


Figure 6. Two patterns of endemism in North America obtained from the matrix with the equal training sensitivity and specificity threshold. Gray quadrats: Western pattern; black quadrats: Eastern pattern.

DISCUSSION

It is known that the species distribution models have limitations when there are few numbers of occurrences (less than 5)^{18,20,33}. The performance of our models, in terms of AUC, however, did not show any differences with few and many records. None of the species had a value lower than 0.7 of AUC for training and testing. This can be due to the fact that Maxent performs well with small samples of records¹⁸; although it can be due also to some intrinsic feature of AUC, because the increment to geographical extents outside presence environmental domain generates higher scores of AUC⁴⁵.

Most species had values lower than 0.1 for the minimum training presence; whilst most mammals had values around 0.5 for the tenth percentile presence and 0.3 for the equal training sensitivity and specificity. Because our data came from museum collections in databases and bibliography, and despite our geographic validation, it is possible that some of them have outliers represented by inconsistencies in georeference or identification of species, even after our verification. Then, those outliers can affect the minimum training presence lower value, because it forces the threshold to include them. However, it is possible that the minimum training presence threshold can be used when the input data had undergone a strict identification of outliers previous to the modelling, or when the data are from very systematic fieldwork, as in Giovanelli *et al.*⁴¹

We found that the more consistent identification of areas of endemism was obtained using the tenth percentile training presence threshold, followed by the 0.5 presence probability, at the same level to the equal training sensitivity and specificity, and the worst for the minimum training presence. The latter resulted the worst threshold, because it tends to enlarge too much the areas of distribution of the taxa, specially in cases where data come from several sources and dissimilar sample effort. Moreover some points can be out of the range of distribution of the modeled species (outliers), because recent taxonomic or nomenclatural changes. Again, it can be relevant to perform an analysis of identification of outliers before the modelling. According to our results, the best option is to use the tenth percentile training presence, which considers the probability at which 10% of the training presence records are omitted, specially the outliers. Other authors have used successfully the 20th percentile in order to avoid bias by outlying records⁴⁰.

The 0.5 presence probability threshold can be a good statistical option and a standard measure for all taxa, but it should be used cautiously, because it may under-identify some areas of endemism. Although some authors suggest that a threshold fixed a priori yields a binary model that is not biologically meaningful and not necessarily results in high accuracy^{16,17}, as 0.5, our study support the statement that this threshold is more restrictive than a lowest presence threshold. Waltari & Guralnick³⁵

mentioned that the 0.5 (50) threshold identified smaller areas than the lowest presence threshold, and we agree with them. They also mentioned that the latter may include population sinks not located in long-term suitable areas. So, they proposed that the 0.5 threshold can be underpredicting habitat suitability, however, we think that this does not necessarily occur. These authors chose both thresholds (conservative and restricted), because the potential distribution at the threshold chosen only represents the widest possible extent of a species.

Pearson *et al.*¹⁸ selected two thresholds: the lowest presence threshold, being conservative and identifying the minimum predicted area possible whilst maintaining zero omission error in the training data; and a more liberal fixed thresholds that rejected only the lowest 10% of possible predicted values. Papes & Gaubert³³, following Pearson *et al.*¹⁸, mentioned that the acceptable threshold value will depend of the question: if the interest are general patterns, the liberal threshold is suitable, but for conservation where the over-prediction is not desirable, the conservative threshold is more adequate. For the identification of areas of endemism, we consider that it is necessary to use a conservative threshold, because a liberal threshold tends to mask some patterns. For example, the Nearctic pattern cannot be recovered, although there are five species that share their distributions²⁷. It is surprising that the Northern pattern was not recovered with any threshold. It was originally discovered with three endemic species²⁷, although the overlapping of their distributional areas is evident, but the models show a discontinuity (at central Canada) that may affect the identification of the area of endemism.

Pearson *et al.*¹⁸ also found that it is possible to use a threshold lower than the lowest presence threshold (threshold 10, equivalent to our 0.1) when small numbers of presence data are available. In our case, it was not necessary, because even the tenth percentile training presence was better than the minimum training presence, and a lower threshold will prevent the correct identification of areas of endemism.

CONCLUSIONS

The identification of areas of endemism represents one of the main goals in biogeography. Its accurate identification depends on the appropriate inference of the individual areas of distribution. Although the field of selection of thresholds in modelling potential distributions is yet controversial, it is possible to obtain better results in analysis of endemism using the best approximation to real distributional areas. The testing of several thresholds before analyzing areas of endemism could be relevant in the identification of distributional patterns of the taxa, however, a threshold similar to the tenth percentile training presence can offer good results.

ACKNOWLEDGEMENTS

Niza Gámez, Rode A. Luna, Ana Lilia González, Estela Rivera and Lucero Cetina helped us with the integration of the database and the generation of the models. We thank the support of CONACyT project 80370. We thank the commentaries from Sergio Roig-Juñent, Patricio Plissock and Juan J. Morrone.

REFERENCES

- Ramírez-Barahona, S., Torres-Miranda, A., Palacios-Ríos, M. & Luna-Vega, I. Historical biogeography of the Yucatan Peninsula, Mexico: a perspective from ferns (Monilophyta) and lycopods (Lycophyta). *Biol. J. Linn. Soc.* **98**, 775-786 (2009).
- Espadas-Manrique, C., Durán, R. & Argáez, J. Phytogeographic analysis of taxa endemic to the Yucatan Peninsula using geographic information systems, the domain heuristic method and parsimony analysis of endemism. *Divers. Distrib.* **9**, 313-330 (2003).
- Rojas-Soto, O.R., Alcántara-Ayala, O. & Navarro, A.G. Regionalization of the avifauna of the Baja California Peninsula, Mexico: A parsimony analysis of endemism and distributional modeling approach. *J. Biogeogr.* **30**, 449-461 (2003).
- Escalante, T., Sánchez-Cordero, V., Morrone, J.J. & Linaje, M. Areas of endemism of Mexican terrestrial mammals: A case study using species' ecological niche modeling, Parsimony Analysis of Endemism and Goloboff fit. *Interciencia* **32**, 151-159 (2007).
- Escalante, T. *et al.* Ecological niche models and patterns of richness and endemism of the southern Andean genus *Eurymetopum* (Coleoptera: Cleridae). *Rev. Bras. Entomol.* **53**, 379-385 (2009).
- Escalante, T., Szumik, C. & Morrone, J.J. Areas of endemism of Mexican mammals: Re-analysis applying the optimality criterion. *Biol. J. Linn. Soc.* **98**, 468-478 (2009).
- Pearson, R.P. *et al.* Model-based uncertainty in species range prediction. *J. Biogeogr.* **33**, 1704-1711 (2006).
- Elith, J. *et al.* Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **29**, 129-151 (2006).
- Plissock, P. & Fuentes-Castillo, T. Modelación de la distribución de especies y ecosistemas en el tiempo y en el espacio: una revisión de las nuevas herramientas y enfoques disponibles. *Rev. Geogr. Norte Gd.* **48**, 61-79 (2011).
- Phillips, S.J., Anderson, R.P. & Schapire, R.E. A maximum entropy modelling of species geographic distributions. *Ecol. Model.* **190**, 231-259 (2006).
- Phillips, S.J. & Dudík, M. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography* **31**, 161-175 (2008).
- Elith, J. *et al.* A statistical explanation of MaxEnt for ecologists. *Divers. Distrib.* **17**, 43-57 (2011).
- Liu, C., Berry, M., Dawson, T.P. & Pearson, R.G. Selecting thresholds of occurrence in the prediction of species distributions. *Ecography* **28**, 385-393 (2005).
- Pawar, S. *et al.* Conservation assessment and prioritization of areas in Northeast India: Priorities for amphibians and reptiles. *Biol. Conserv.* **136**, 346-361 (2007).
- Manel, S., Williams, H.C. & Omerod, D.J. Evaluating presence-absence models in ecology: The need to account for prevalence. *J. Appl. Ecol.* **38**, 921-931 (2001).
- Jiménez-Valverde, A. & Lobo, J.M. Threshold criteria for conversion of probability of species presence to either-or presence-absence. *Acta Oecol.* **31**, 361-369 (2007).
- Freeman, E.A. & Moisen, G.G. A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecol. Model.* **217**, 48-58 (2008).
- Pearson, R.G., Raxworthy, C.J., Nakamura, M. & Peterson, T. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *J. Biogeogr.* **34**, 102-117 (2007).
- Aranda, S.D. & Lobo, J.M. How well does presence-only-based species distribution modelling predict assemblage diversity? A case study of the Tenerife flora. *Ecography* **34**, 31-38 (2011).
- Bean, W.T., Stafford, R. & Brashares, J.S. The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution model. *Ecography* **35**, 250-258 (2012).
- Morrone, J.J. *Evolutionary biogeography: An integrative approach with case studies* (Columbia University Press, New York, 2009). 301 pp.
- Morrone, J.J. On the identification of areas of endemism. *Syst. Biol.* **43**, 438-441. (1994).
- Escalante, T. Un ensayo sobre regionalización biogeográfica. *Rev. Mex. Biodivers.* **80**, 551-560 (2009).
- Szumik, C.A., Cuezco, F., Goloboff, P.A. & Chalup, A.E. An optimality criterion to determine areas of endemism. *Syst. Biol.* **51**, 806-816 (2002).
- Szumik, C.A. & Goloboff, P.A. Areas of endemism: An improved optimality criterion. *Syst. Biol.* **53**, 968-977 (2004).
- Estrada, Y.-Q., Luna, R.A. & Escalante, T. Patrones de distribución de los mamíferos en la provincia Oaxaca-Tehuacanense, México. *Therya* **3**, 33-51 (2012).
- Escalante, T., Rodríguez-Tapia, G., Szumik, C., Morrone, J.J. & Rivas, M. Delimitation of the Nearctic region according to mammalian distributional patterns. *J. Mammal.* **91**, 1381-1388 (2010).
- Arita, H.T. & Rodríguez, G. Patrones geográficos de diversidad de los mamíferos terrestres de América del Norte. Instituto de Ecología, UNAM. SNIB-Conabio database, project Q068 (2004).
- ESRI. ArcGis v. 9.3. Redlands, CA. (2009).
- Hall, E.R. *The mammals of North America. Vols. I and II* (John Wiley and Sons, New York, 1981). 1181 pp.
- Ceballos, G. & Oliva, G. *Los mamíferos silvestres de México* (Comisión Nacional para el Conocimiento y Uso de la Biodiversidad - Fondo de Cultura Económica, México, D.F., 2005). 986 pp.
- Hijmans, R.J., Cameron, S. & Parra, J. *WorldClim v. 1.3*. University of California, Berkeley (<http://biogeog.berkeley.edu/worldclim/worldclim.htm>) (2005).
- Papes, M. & Gaubert, P. Modelling ecological niches from low numbers of occurrences: assessment of the conservation status of poorly known viverrids (Mammalia, Carnivora) across two continents. *Divers. Distrib.* **13**, 890-902 (2007).
- Loiselle, B.A. *et al.* Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes. *J. Biogeogr.* **35**, 105-116 (2008).
- Waltari, E. & Guralnick, R.P. Ecological niche modeling of montane mammals in the Great Basin, North America: examining past and present connectivity of species across basins and ranges. *J. Biogeogr.* **36**, 148-161 (2009).

36. Costa, G.C., Nogueira, C., Machado, R.B. & Colli, G.R. Sampling bias and the use of ecological niche modeling in conservation planning: A field evaluation in a biodiversity hotspot. *Biodivers. Conserv.* **19**, 883-899 (2009).
37. Brito, J.C., Acosta, A.L., Álvares, F. & Cuzin, F. Biogeography and conservation of taxa from remote regions: An application of ecological-niche based models and GIS to North-African canids. *Biol. Conserv.* **142**, 3020-3029 (2009).
38. Newbold, T., Gilbert, F., Zalat, S., El-Gabbas, A. & Reader, T. Climate-based models of spatial patterns of species richness in Egypt's butterfly and mammal fauna. *J. Biogeogr.* **36**, 2085-2095 (2009).
39. Colacicco-Mayhugh, M.G., Masuoka, P.M. & Grieco, J.P. Ecological niche model of *Phlebotomus alexandri* and *P. papatasi* (Diptera: Psychodidae) in the Middle East. *Int. J. Health Geogr.* **9**, 2-9 (2010).
40. Donegan, T.M. & Avendaño, J.E. A new subspecies of mountain tanager in the *Anisognathus lacrymosus* complex from the Yariguies Mountains of Colombia. *Bull. BOC* **130**, 13-32 (2010).
41. Giovanelli, J.G.R., Ferreira de Siqueira, M., Haddad, C.F.B. & Alexandrino, J. Modeling a spatially restricted distribution in the Neotropics: How the size of calibration area affects the performance of five presence-only methods. *Ecol. Model.* **221**, 215-224 (2010).
42. Torres, R. & Jayat, J.P. Modelos predictivos de distribución para cuatro especies de mamíferos (Cingulata, Artiodactyla y Rodentia) típicas del Chaco en Argentina. *Mastozoología Neotropical* **17**, 335-352 (2010).
43. Fielding, A.H. & Bell, J.F. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environ. Conserv.* **24**, 38-49 (1997).
44. Goloboff, P.A. Programs for identification of areas of endemism. <http://www.zmuk.dk/public/phylogeny/endemism> (2005).
45. Lobo, J.M., Jiménez-Valverde, A. & Real, R. AUC: A misleading measure of the performance of predictive distribution models. *Global Ecol. Biogeogr.* **17**, 145-151 (2008).