Biodivers Conserv
DOI 10.1007/s10531-009-9746-8

ORIGINAL PAPER

Sampling bias and the use of ecological niche modeling in conservation planning: a field evaluation in a biodiversity hotspot

Gabriel C. Costa · Cristiano Nogueira · Ricardo B. Machado · Guarino R. Colli

Received: 19 June 2009 / Accepted: 29 October 2009

Abstract Ecological niche modeling (ENM) has become an important tool in conservation biology. Despite its recent success, several basic issues related to algorithm performance are still being debated. We assess the ability of two of the most popular algorithms, GARP and Maxent, to predict distributions when sampling is geographically biased. We use an extensive data set collected in the Brazilian Cerrado, a biodiversity hotspot in South America. We found that both algorithms give richness predictions that are very similar to other traditionally used richness estimators. Also, both algorithms correctly predicted the presence of most species collected during fieldwork, and failed to predict species collected only in very few cases (usually species with very few known localities, i.e., <5). We also found that Maxent tends to be more sensitive to sampling bias than GARP. However, Maxent performs better when sampling is poor (e.g., low number of data points). Our results indicates that ENM, even when provided with limited and geographically biased localities, is a very useful technique to estimate richness and composition of unsampled areas. We conclude that data generated by ENM maximize the utility of existing biodiversity data, providing a very useful first evaluation. However, for reliable conservation decisions ENM data must be followed by well-designed field inventories, especially for the detection of restricted range, rare species.

G. C. Costa (✉)
Centro de Biociências, Departamento de Botânica, Ecologia e Zoologia, Universidade Federal do Rio Grande do Norte, Campus Universitário, Lagoa Nova, 59072-970 Natal, RN, Brazil
e-mail: costagc@cb.ufrn.br; costagc@mac.com

C. Nogueira · R. B. Machado · G. R. Colli
Departamento de Zoologia, Universidade de Brasília, CEP: 70910-900 Brasília, DF, Brazil
e-mail: cnogueira@unb.br

R. B. Machado
e-mail: rbmac@unb.br

G. R. Colli
e-mail: grcolli@unb.br

Published online: 11 November 2009

Keywords Biodiversity hotspots · Brazil · Cerrado · Conservation planning · Ecological niche modeling · GARP · Maxent · Sampling bias · Species distribution · Squamates

Introduction

Sound conservation strategies depend heavily on biodiversity information, especially species distributions. However, knowledge about biodiversity remains incomplete, particularly in the highly speciose Tropics, where many species remain formally undescribed (Linnean shortfall) and poorly understood in terms of their geographical distribution (Wallacean shortfall) (Lomolino 2004; Whittaker et al. 2005). As a result, biodiversity databases, although extremely useful, may suffer strong limitations even for groups and/or regions that have been well studied (Soberón et al. 2000; Hortal et al. 2007; Soberón et al. 2007). Recently a new methodological approach, ecological niche modeling (ENM), has emerged as a powerful tool to reconstruct or predict species distributions. The method uses geo-referenced known occurrence points of the species under study that are linked with abiotic and/or biotic variables from each point locality. A particular algorithm processes information and then a predicted 'niche' envelope in which the species is likely to occur is produced (see Elith et al. 2006 for a review of the methods).

Ecological niche modeling has been applied in conservation biology to identify areas with high species richness (Garcia 2006; Costa et al. 2007), sample for rare species (Guisan et al. 2006), predict effects of climate change on species' distributions (Araújo and Rahbek 2006; Hijmans and Graham 2006), and assess potential invasion and proliferation of exotic species (Peterson and Vieglais 2001). Despite the recent growth and diversity of studies that apply ENM to address conservation and/or evolutionary questions, several basic issues related to the performance of the algorithms remain unsettled. Among the most important issues is how the accuracy of ENM is influenced by factors such as sample size (Stockwell and Peterson 2002; Hernandez et al. 2006), spatial scale (Lassueur et al. 2006; Guisan et al. 2007a; Trivedi et al. 2008), the nature of the environmental data set (Parra et al. 2004; Peterson and Nakazawa 2008), species traits (Poyry et al. 2008), biotic interactions (Araújo and Luoto 2007; Heikkinen et al. 2007), and finally, which particular algorithm is being used (Segurado and Araújo 2004; Elith et al. 2006).

Another important issue is how ENM models are influenced by geographical bias in the sampling points used to train the models. For example, a previous study found that the frequency of plant observations near roads was greater than that expected from a spatially random distribution, such that predictive maps based on near-road observations were less accurate than those based on observations corrected for roadside bias (Kadmon et al. 2004). On a larger spatial scale, Loiselle et al. (2008) found that although localities based on herbarium collections did not represent well the entire climatic gradient in which most species occur, this existing climatic bias did not greatly affect distribution predictions when compared with an unbiased data set. Therefore, determining how well ENM is able to reconstruct the entire distribution of a species when input data comes from only a biased subset of the whole species range is a crucial matter to establish ENM utility as a conservation tool.

We use two of the most commonly used ENM algorithms (GARP and Maxent) to predict the distribution of squamate reptiles (lizards, snakes, and amphisbaenians) in the Brazilian Cerrado, one of the 34 world biodiversity hotspots (Myers 2003; Mittermeier

et al. 2005), a region for which a strong sampling bias exists (Costa et al. 2007). We tested the performance of these two methods by first predicting species richness and composition of an unsampled area of conservation interest using ENM, and then conducting field surveys to determine actual species richness and composition. The poorly sampled region lies near the northern edge of the Cerrado region. We identify both limitations and strengths of ENM as a tool in conservation planning and biodiversity studies.

Methods

Ecological niche modeling

We used GARP and Maxent to model the distributions of all known (at the time of analysis) squamate species occurring in the Cerrado, a total of 237 species based on an extensive existing database. We used only species for which at least one data point existed within the Cerrado region. For species whose distributions spanned multiple biomes, we also included data points outside of the Cerrado, because characteristics of these points can help identify suitable regions for species occurrence within Cerrado. Locality data for each species were collected from museums, literature (only taxonomic revisions or voucher based lists), and previous fieldwork (see Costa et al. 2007; Nogueira et al. 2009). All museum specimen records were checked for accurate taxonomy and the most precise locality information, a critical need, as museum data can be error-prone. Thus, we did not include in our analyses records obtained in electronic databases, that often include unchecked, raw museum data. Locality data varied between 3 and 256 (mean = 35.58, standard deviation = 39.32) unique point localities per species. The dataset contains a clear geographical sampling bias; most records come from the central and southeastern portion of the Cerrado, where the majority of research institutions are located, and very few inventories have been made in the Northern parts of the Cerrado region (Fig. 1).

We used the implementation of GARP provided by the software OpenModeller. The algorithm divides occurrence points into training and extrinsic test data. The extrinsic test dataset is divided evenly into true training data (for model rule development) and intrinsic test data (for model rule evaluation and refinement). Models are based on presence-only data, with absence information included via random sampling of 1,250 pseudo-absence points from the set of pixels at which the individual species were not collected. The algorithm works in an iterative process of rule selection, testing, and incorporation or rejection. More details on algorithm function are provided by Stockwell and Noble (1992). We used the default parameters of the OpenModeller version of GARP with the best subset selection option (optimum models considering omission/commission relationships; see Anderson et al. 2003).

Maxent fits a probability distribution for species occurrence to the set of pixels across the region of interest. The algorithm is based on the principle that, given the appropriate constraints, the best explanation to unknown phenomena will maximize the entropy of the probability distribution. For ecological niche modeling, these constraints derive from the values of those pixels at which the species has been detected. More details on Maxent function are provided by Phillips et al. (2004, 2006). We used the default parameters for Maxent v.3.2.1, which were adjusted based on a recent comprehensive evaluation (Phillips and Dudik 2008). The output format for Maxent and GARP are raster grids with values ranging from 0–1 for Maxent and 0–100 for GARP. To transform the models into discrete presence or absence, selection of a threshold is necessary. To select an appropriate

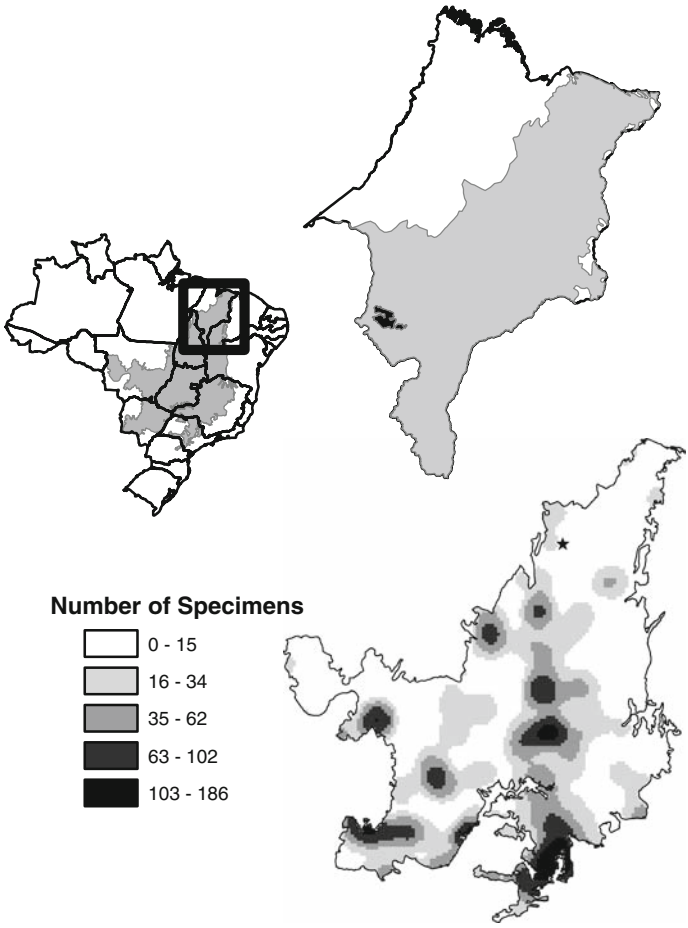

Fig. 1 Map of the study area and sampling profile for the dataset for the Brazilian Cerrado Squamates. On the upper figure, *gray shade* corresponds to the limits of the Cerrado region. In detail, the state of Maranhão where PNCM (limits are shown in *black*) is located. On the lower figure, Kernel density function was applied using all sampling points to create a smooth tapered surface. Darker regions indicate higher density of sampling points (more records are available on those regions). *Star symbol* represent the location of our field site in the northern portion of Cerrado

threshold we first defined a user-selected parameter *E*, which refers to the amount of error associated with the presence localities dataset (see Peterson et al. 2008 for details). Although we checked specimens identities from the collection databases, we georeferenced most of the localities ourselves based on localities descriptions from the museum records. Therefore, there could be some error associated with our dataset, so we set *E* to 5%. Next, we determined the lowest predicted value associated with any one of the observed presence records [i.e., lowest presence threshold 'LPT' (Pearson et al. 2007)]. We then set our threshold at LPT–*E* (i.e., from the distribution of predicted values associated with presence records, we eliminated the lowest 5% and set our threshold at the remaining lowest value).

For both GARP and Maxent, we used environmental variables from the Worldclim project (Hijmans et al. 2005) and a Normalized Difference Vegetation Index layer (NDVI),

based on the average values of satellite images from the end of the dry season and the end of the wet season. We constructed a correlation matrix among all variables and selected for the modeling only variables that were not highly correlated ($r > 0.9$). After applying this criterion, we used the following environmental variables: altitude, annual precipitation, isothermality, maximum temperature of warmest month, mean diurnal range, mean temperature of warmest quarter, mean temperature of wettest quarter, minimum temperature of coldest month, precipitation of coldest quarter, precipitation of driest month, precipitation of warmest quarter, precipitation of wettest month, precipitation seasonality, temperature annual range, temperature seasonality, and NDVI. All variables were at 1 km resolution.

Model evaluations

To statistically evaluate model performance we used a recently proposed modification of the area under the curve—AUC on receiver operating characteristic—ROC, named the partial ROC approach (Peterson et al. 2008). ROC analysis is a method designed to evaluate the specificity (absence of commission error) and sensitivity (absence of omission error) of a diagnostic test (Zweig and Campbell 1993; Fielding and Bell 1997). The AUC provides a threshold-independent measure of model performance as compared with that of null expectations (Fielding and Bell 1997), and is the most commonly used statistic to evaluate ENM performance (Elith et al. 2006; Guisan et al. 2007b; Peterson et al. 2007). The partial ROC procedure is particularly suitable for our situation because we are comparing the performance of methods that do not provide predictions across the same spectrum of proportional areas in the study area. Partial AUC values are presented as a ratio between the AUC (with modified x-axis, from traditional applications) and the null expectation of AUC (which unlike traditional ROC approaches is not equal 0.5, and is variable). For a detailed account of the partial ROC approach please refer to Peterson et al. (2008).

After constructing niche models and calculating the partial ROC statistics, we tested performance of ENM in predicting species diversity and distributions by surveying a remote and previously unsampled area. This allowed us to evaluate the effect of sampling bias on the ability of ENM to project distributions into unsampled regions, and to determine whether GARP and/or Maxent are differentially affected by sampling bias. Using this approach, several scenarios are possible. First, when sampling points are concentrated in a subset of the species range, ENM is (a) capable of predicting the species occurrence or (b) not able to predict the occurrence of the species in the unsampled region outside of the major concentration of sampling (Fig. 2a, b). Second, when the sampling points are more dispersed throughout a species' range, ENM is (c) capable of predicting the species occurrence or (d) not able to predict the species occurrence (Fig. 2c, d) in the area.

Study area and field sampling

We chose a study site located within the northern portion of the Cerrado region in the "Parque Nacional da Chapada das Mesas" (PNCM—7°10′S, 47°9′W), a recently created 160,000 ha conservation unit in the Brazilian state of Maranhão (Fig. 1). This area is ideal for evaluating sampling bias in ENM because it is relatively undisturbed, poorly sampled, and a recent niche modeling exercise predicted high squamate species diversity there (Costa et al. 2007).

We collected squamates from November 30th to December 17th 2007, using 48 arrays of pitfall traps and 24 arrays of funnel traps resulting in 5,184 trap*days. Traps were

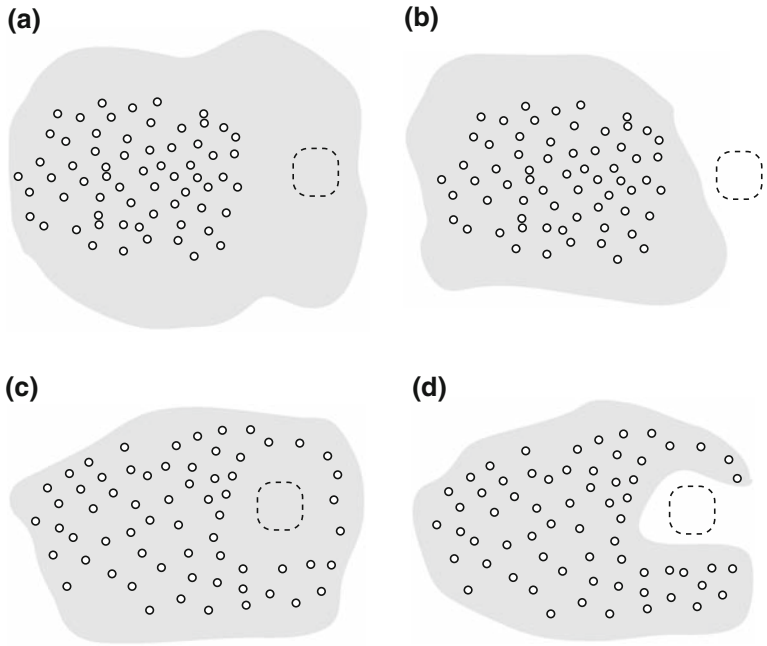

Fig. 2 Diagram representing different possible scenarios for ENM in different biased sampling scenarios. *Circles* represent known occurrence localities, *dashed line* represents area being surveyed, and *gray area* represents the predicted distribution of the species based on ENM. See text for details

divided among six sampling points, which were located inside PNCM and were chosen in order to sample the full range of landscape and vegetation cover variation within PNCM. Each array of pitfall traps consisted of four 35 l buckets arranged in a Y-shape (one at the center and one on each of the three ends). Buckets were 5 m from each other, and 50 cm high plastic fences (bottom edge buried) spanned the distance between buckets. The funnel trap arrays consisted of a single 5 m long, 50 cm high plastic fence with a pair of funnel traps at each end (one on each side). Arrays were spaced approximately 20 m apart. All traps were checked daily. All specimens collected were deposited at the Coleção Herpetológica da Universidade de Brasília (CHUNB). In addition to our trapping methods, we collected animals by hand, noose, or using a shotgun during haphazard searches of various habitats within PNCM. We also routinely drove roads both during the day and night for snakes in the process of crossing or that had been killed by vehicles. Road collecting is a common and effective survey method for snakes (Sullivan 1981).

To estimate species richness of the region based on our sampling, we produced a species accumulation curve using the software EstimateS v.8 (Colwell 2005). EstimateS randomizes the sampling order to generate smooth species accumulation curves and species richness estimators. Resulting values are numbers of species expected based on empirical data (Colwell et al. 2004). We used the average of several richness estimators provided by the software (ACE, ICE, Chao1, Chao2, Jack 1, Jack 2, Bootstrap, MMruns, Cole) to estimate species richness based on the sampling (Chazdon et al. 1998; Chao et al. 2000), and performed 10,000 randomizations without replacement. In addition, we fitted our data to three different accumulation curve mathematical models, Clench, Logarithmic, and Exponential. Model fitting was performed using methods and software described by Díaz-Francés and Soberón

(2005). The model providing the best fit can then be used to estimate the asymptote (i.e., total species richness) of the species accumulation curve.

Statistical analysis

We used the statistical package R (R Development Core Team 2008) to perform a two-sample test for equality of proportions with continuity correction to determine whether a difference exists in the proportion of species successfully predicted between GARP and Maxent. We also developed a multiple logistic regression model to explore different factors that may influence the probability of GARP and Maxent to successfully predict species occurrence in our study site. The dependent variable was the prediction success (0 = fail, 1 = success), and our independent variables were: 1—nearest neighbor index, calculated based on the average distance of each point to its nearest point. Low values of the index indicate a distribution more clumped than expected by chance whereas high values indicate a more dispersed distribution; 2—number of locality points used in the modeling exercise; and 3—distance from the nearest locality point to PNCM. We ran the regression with all species ($N = 48$) and also, to investigate the influence of species with low number of sampling points, we ran the regression using only species with more than 15 known locality points ($N = 42$).

For both methods, a species was considered present in PNCM if any pixel of the final predicted distribution map (see above for details on how we obtained the final presence/absence maps) for that species lied within the PNCM limits (Fig. 1). Because we cannot distinguish between species that do not occur in the region from species that do occur but were not collected because of sampling deficiency, we restricted our evaluations only to species collected during the field survey. One species (*Amphisbaena* sp.) was removed from all analyses due to taxonomic uncertainties. Calculations of the nearest neighbor index and distance to the nearest point were performed in ArcGIS 9.2. The multiple logistic regression was performed in R (R Development Core Team 2008).

Results

We collected a total of 49 species of squamates in PNCM (Table 1). Our accumulation curve analysis indicated that our sampling efforts were far from stabilizing and the true richness of squamates in the region may be over 70 species (Fig. 3). The average of richness estimators provided by EstimateS was 73.35 species and the logarithmic model produced the best fit. Usually, when this model produces the best fit it is because the sample area is too large and/or the taxa are poorly known (Soberón and Llorente 1993). Such results are well known for Neotropical squamates, which require long-term fieldwork for sampling to stabilize (Duellman 1978), often because of snake species that are rare or difficult to sample.

GARP predicted 72 species within PNCM; we collected 43. The method failed to predict the presence of five species collected in our survey. Maxent predicted 74 species within PNCM and, among those, we collected 39. Maxent failed to predict the presence of nine species we collected in our fieldwork (Table 1). In addition, there was no difference in the ratio between predicted and surveyed species between GARP and Maxent ($\chi^2 = 0.47$, $P = 0.49$). Maxent models had higher Partial AUC values (GARP $\bar{x} = 1.41 \pm 0.20$, median = 1.37; Maxent $\bar{x} = 1.59 \pm 0.12$, median = 1.60; $F = 26.73$, $P < 0.01$; all partial AUC values are in Table 1).

Table 1 Species collected in the survey of PNCM

Species	Number of points	Garp	Maxent
Ameiva ameiva	249	1 (1.10)	1 (1.25)
Amphisbaena alba	41	1 (1.61)	1 (1.59)
Anolis chrysolepis	46	1 (1.30)	0 (1.43)
Apostolepis cearensis	3	0 (1.40)	0 (1.99)
A. polylepis	3	0 (1.29)	0 (1.99)
Boa constrictor	111	1 (1.22)	1 (1.63)
Bothrops lutzi	15	1 (1.27)	1 (1.83)
B. moojeni	112	1 (1.42)	0 (1.70)
Chironius exoletus	21	1 (1.38)	1 (1.55)
C. flavolineatus	38	1 (1.33)	1 (1.51)
Cnemidophorus mumbuca	4	0 (1.25)	0 (1.95)
Colobosaura modesta	44	1 (1.58)	1 (1.67)
Corallus hortulanus	16	1 (1.49)	1 (1.78)
Drymarchon corais	55	1 (1.40)	1 (1.50)
Epicrates cenchria	103	1 (1.26)	1 (1.54)
Gymnodactylus carvalhoi	36	1 (1.37)	1 (1.63)
Hemidactylus mabouia	110	1 (1.13)	1 (1.40)
Hydrodynastes bicinctus	27	1 (1.64)	1 (1.60)
Iguana iguana	100	1 (1.13)	1 (1.48)
Imantodes cenchoa	24	1 (1.35)	1 (1.49)
Kentropyx calcarata	85	1 (1.21)	1 (1.52)
Leptotyphlops brasiliensis	5	0 (1.0)	1 (1.93)
Liophis almadensis	70	1 (1.35)	1 (1.52)
L. poecilogyrus	185	1 (1.30)	1 (1.66)
L. reginae	82	1 (1.44)	1 (1.60)
Liotyphlops ternetzii	9	0 (1.19)	0 (1.75)
Mabuya heathi	54	1 (1.46)	0 (1.64)
Mabuya nigropunctata	132	1 (1.20)	1 (1.42)
Mastigodryas bifossatus	104	1 (1.38)	1 (1.63)
Micrablepharus maximiliani	50	1 (1.40)	0 (1.72)
Micrurus brasiliensis	9	1 (1.58)	0 (1.86)
Oxyrhopus trigeminus	111	1 (1.77)	1 (1.61)
Philodryas nattereri	94	1 (1.38)	1 (1.65)
P. olfersi	90	1 (1.27)	1 (1.56)
Phimophis guerini	38	1 (1.91)	1 (1.62)
P. iglesiasi	9	1 (1.98)	1 (1.76)
Pseudoboa neuwiedii	24	1 (1.53)	1 (1.61)
P. nigra	60	1 (1.41)	1 (1.56)
Psomophis joberti	46	1 (1.87)	1 (1.66)
Sibynomorphus mikanii	97	1 (1.52)	1 (1.71)
Spilotes pullatus	67	1 (1.28)	1 (1.52)
Tantilla melanocephala	37	1 (1.25)	1 (1.49)
Thamnodynastes hypoconia	21	1 (1.51)	1 (1.81)

Table 1 continued

Species	Number of points	Garp	Maxent
Tropidurus oreadicus	50	1 (1.30)	1 (1.56)
Tupinambis merianae	66	1 (1.30)	1 (1.44)
T. teguixin	48	1 (1.29)	1 (1.41)
Typhlops brongersmianus	24	1 (1.63)	1 (1.68)
Waglerophis merremi	136	1 (1.22)	1 (1.62)

Number of points used to train the model, GARP and Maxent predictions (0 = did not predict, 1 = predict to occur at PCNM). Results of the partial ROC analysis for both methods are shown within parenthesis (mean across 200 bootstrap replicates)

Amphisbaena sp. was collected but not used in the analysis due to taxonomic uncertainties

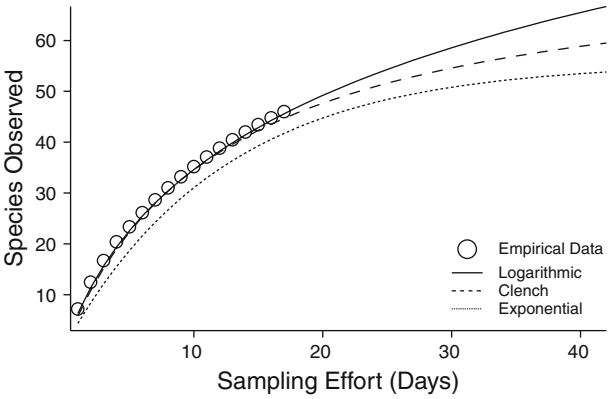

Fig. 3 Results of the accumulation curve analysis. *Open circles* represent mean values from 10,000 randomizations without replacement of the original matrix

GARP predicted four species we collected in the area that Maxent failed to predict. Among those, three species followed the pattern described in Fig. 2a, where the known sampled localities were concentrated in the central part of the Cerrado. None or very few known localities were in the northern part of the Cerrado where PNCM is located. We illustrate three of those cases in Fig. 4a, b, and d. The remaining species follow a pattern similar to the one described in Fig. 2c, where sampling is more spread throughout the Cerrado and PNCM was surrounded by a few known sampled localities. We illustrate this case in Fig. 4c.

Maxent successfully predicted one species in PNCM that GARP failed to predict (*Leptotyphlops brasiliensis* Fig. 5a); this species had low numbers of known localities (<10 know localities). Both methods successfully predicted 38 species and failed to predict four species we actually collected in the area. With the exception of *Liotyphlops ternetzii* (Fig. 5c), all other species that both methods failed to predict were likely affected by the very low numbers of known localities available for the modeling (<5 know localities, i.e., *Apostolepis polylepis* Fig. 5b).

The multiple logistic regression results show that GARP models were not significantly influenced by any of the independent variables in the regression model. The same result

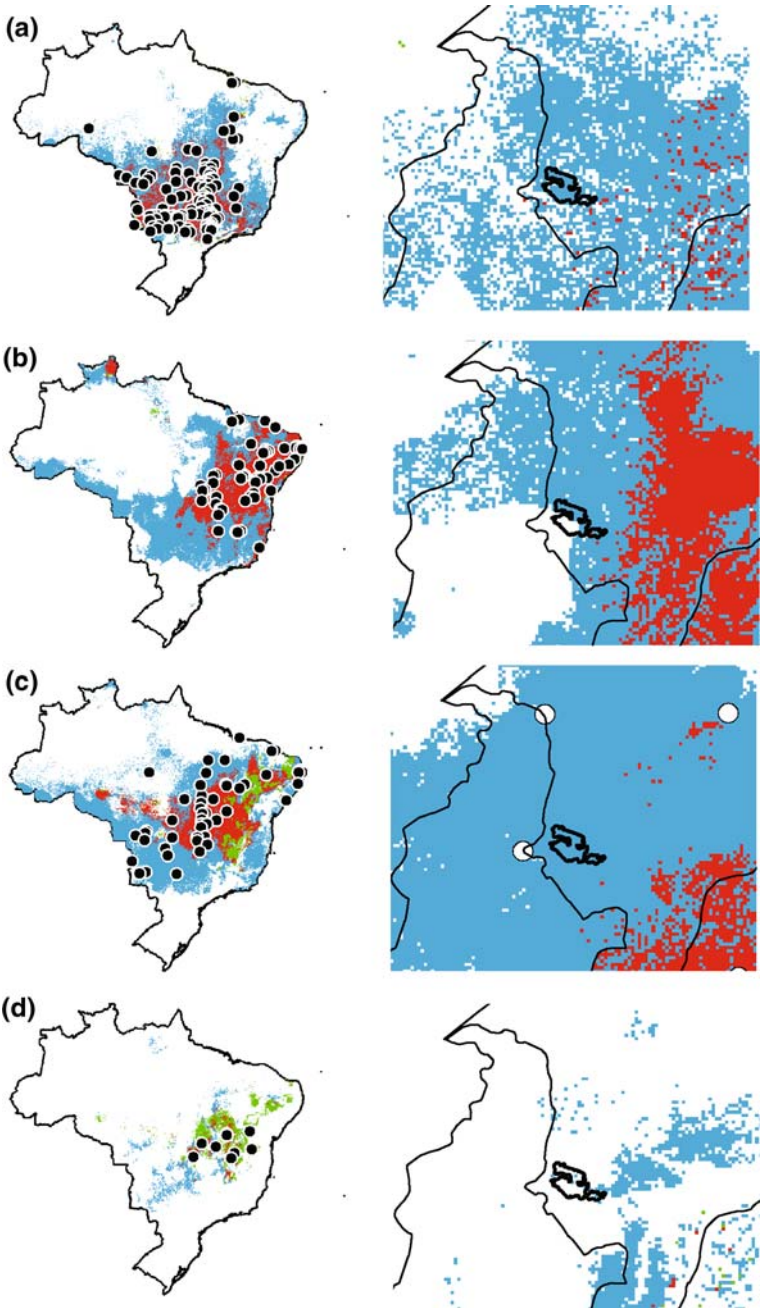

Fig. 4 Example of ENM results where GARP successfully predicted the species to occur in PNCM whereas Maxent failed. *Circles* represents known occurrence localities, *blue* represents GARP predictions, *green* Maxent predictions, and *red* the coincidence of both methods. **a** *Bothrops moojeni*, **b** *Mabuya heathi*, **c** *Micrablepharus maximiliani*, **d** *Micrurus brasiliensis*. (Color figure online)

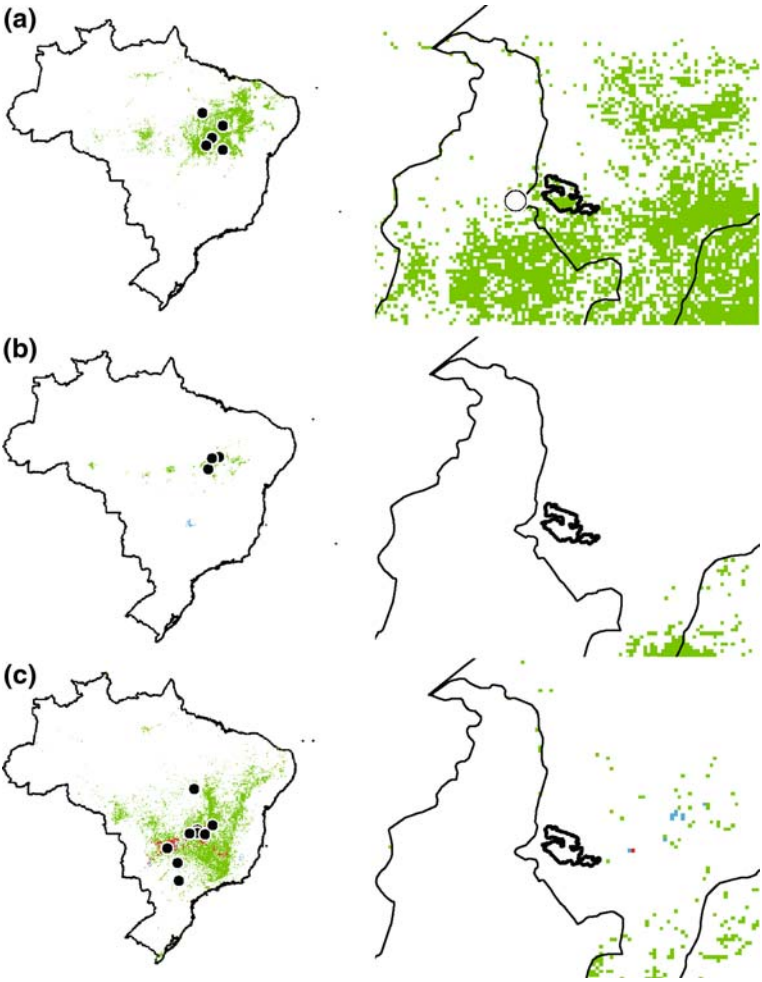

Fig. 5 ENM results where Maxent successfully predicted the species to occur in PNCM whereas GARP failed, **a** *Leptotyphlops brasiliensis*, and examples where both GARP and Maxent failed to predict species occurrence **b** *Apostolepis polylepis*, **c** *Liotyphlops ternetzii*

was found when species with few numbers of known locality points were removed from the analysis (Table 2). Maxent predictions were influenced by the degree of dispersal of sampled points when regression was performed with all species. After eliminating species with few known localities, predictions were influenced only by the distance to the nearest point (Table 2).

Discussion

Based on the accumulation curve analysis, both ENM methods accurately estimated species richness. The richness estimators and accumulation curve predicted that richness for the region should be around 73 species, a rather conservative estimate considering other

Table 2 Results of multiple logistic regression to model the effects of: degree of dispersal of sampled points (nearest neighbor); distance from the closest point (DNP); and number of points used in the modeling, on the ability of GARP and Maxent to successfully predict species occurrence in PNCM (0 = fail, 1 = success)

Predictor	β	SE β	Wald's z	P	e^{β} (odds ratio)
GARP					
Intercept	−45.38/26.57	6963.19/5.28 × 10^5	−0.01/≈0	0.99/≈1	NA
Number of points	27.89/≈0	2185.36/1.53 × 10^3	0.01/≈0	0.99/≈1	1.31 × 10^{12}/1
R_Stat	−140.07/≈0	12216.71/4.25 × 10^5	−0.01/≈0	0.99/≈1	≈0/1
DNP	0.15/≈0	16.33/608.6	0.01/≈0	0.99/≈1	1.16/1
Maxent					
Intercept	−0.29/−14.98	2.26/8.50	−0.13/−1.76	0.90/0.08	NA
Number of Points	0.02/0.04	0.02/0.03	1.52/1.54	0.13/0.12	1.02/1.04
R_Stat	2.24/14.91	1.80/7.59	1.24/1.97	0.21/0.049*	9.33/3 × 10^6
DNP	−0.01/−0.002	<0.01/0.01	−2.11/−0.31	0.04*/0.76	0.99/1

Results after slash are from regression after eliminating species with known locality points lower than 15. Degrees of freedom is equal to 1 in all cases, and sample sizes are 48 and 42. β are the individual regression coefficients, which were tested using Wald's z statistic. e^{β} is the odds ratio, which is the predicted change in odds for a unit increase in the corresponding independent variable. Odds ratios <1 correspond to decreases and odds ratios more than 1.0 correspond to increases in odds. Odds ratios close to 1.0 indicate that unit changes in that independent variable do not affect the dependent variable

* Result is significant at 0.05 significance level

well-sampled Cerrado localities (Colli et al. 2002; França and Araújo 2007; Valdujo et al. 2009). Therefore, our results indicate that predictions from ENM, even when generated by limited and geographically biased dataset, can be a helpful resource to estimate species richness of unsampled regions. This result confirms recent successful applications of ENM in conservation (Domínguez-Domínguez et al. 2006; Garcia 2006; Pawar et al. 2007). It is important to note that, although our input data set is geographically biased towards localities near the major institutions, it still includes point locality information for all 237 species of squamates known to occur in the Cerrado. Therefore, it is possible that a more severe lack of information or stronger bias could generate different results. Thus, although the Cerrado has been traditionally considered one of the most poorly sampled regions in the Neotropics (Colli et al. 2002), an extensive and careful revision and compilation of voucher records in natural history collections and field samplings, coupled with ENM techniques, are now providing new insights on species richness and composition, highlighting the importance of not neglecting available species occurrence information (see discussions in Brooks et al. 2004). Although limited, current knowledge on species distributions must be viewed as the major source of data for illuminating conservation assessments, especially in remote, poorly sampled and highly threatened tropical regions.

Overall, there was no statistical difference on the proportion of correct predictions for PNCM by GARP and Maxent. However, Maxent models produced higher partial AUC values, and GARP predicted more species that were collected on our study area. Two alternative hypotheses may explain why GARP better predict species occurrence in PCNM despite having lower partial AUC models. First, Maxent generally produce better models than GARP, but our approach of evaluating the presence of species in a specific region does not characterize well the performance of models in their entire distribution. In this

scenario, GARP might be overpredicting. For example, consider a model predicting species X to occur in all the Cerrado. It might have a low partial AUC, but will certainly cover the PNCM.

Second, the partial AUC statistics does not provide the best possible evaluation of the models. A lot more data and analyses would be necessary to assess the first hypothesis. However, previous studies have provided support for the second (Raes and ter Steege 2007; Lobo et al. 2008). Some recent work shows that reliance on AUC as the only estimate of model success needs to be re-examined (Austin 2007). Either way, better methods to statistically evaluate ENM models are likely to be a major topic of future research (Raes and ter Steege 2007; Lobo et al. 2008; Peterson et al. 2008).

In some cases, even when most of the known locality points were far away from PNCM, GARP was able to correctly predict species occurrence. This ability of GARP may be desirable in different applications of ENM, including the discovery of new populations and/or species. For example, in Madagascar, field survey of areas with similar sampling characteristics lead researchers to the discovery of several undescribed species of chameleons (Raxworthy et al. 2003). In our system, we discovered no obvious undescribed closely-related species; nevertheless, future genetic and/or morphological studies may reveal hidden diversity because populations of some species appear to be separated by areas where environmental conditions are predicted to be unsuitable. Recent studies in other Cerrado areas are revealing new squamate species, including some which may present restricted ranges, and from poorly studied taxa (Nogueira and Rodrigues 2006; Rodrigues et al. 2007, 2008; Colli et al. 2009). Because these species may show high endemism and restricted ranges, they are of special concern for conservation. Modeling of closely related species may help to identify regions where these species occur, providing useful and unavailable information on biogeographical patterns in the Cerrado.

The ability to project distributions in areas distant from known localities may also be useful in ecosystems such as the Cerrado, where species' range extensions of several hundreds of kilometers are commonly recorded (e.g., Strüssmann and Carvalho 1998; Nogueira 2001; Filho and Montigelli 2006; Freitas et al. 2007; Silveira 2007). This may also be important in other uses of ENM. For some applications of ENM in ecology and evolutionary biology, precisely reconstructing species' distributions is not expected or desired; rather, ENM is used to estimate a map of the environmental space in which the species is likely to occur. In these cases, contrasting where the species is predicted to occur with where the species does occur can provide insights into interesting biogeographical or ecological factors shaping the species' distribution (Anderson et al. 2002; Costa et al. 2008). A method that is too sensitive to sampling bias will be less useful to address such questions. Therefore, the major challenge for ENM is to be able to distinguish models that predict the distribution of the species into areas that are not inhabited (or not sampled) but hold good ecological conditions *versus* models that predict the species to occur in habitats that in fact are not suitable (Peterson 2006; Peterson et al. 2008).

The multiple logistic regression models showed no effect of any variable on GARP ability to successfully predict species' distribution. However, the degree of dispersal of sampled points and the nearest point influenced Maxent. This suggests that, as long as a known locality exists close to the region, the algorithm will successfully predict species' presence even if the distribution of points is clustered. This result also suggests that GARP has a better ability to reconstruct species distribution provided only with a subset of the species known distribution. A recent study found similar results by manipulating a species localities dataset (Peterson et al. 2007).

We found that Maxent produced better estimates of species' distributions than GARP when few localities are used in modeling. This result agrees with a recent study using geckos in Madagascar, which showed Maxent performing better than GARP when sample size was smaller than 10 points (Pearson et al. 2007), and highlights the importance of complementing niche models with detailed and well-designed field inventories. Improving the understanding of distribution patterns for rare or restricted-range species is a major challenge for biogeography and conservation, as these species are dually important: narrow endemics are intrinsically vulnerable to human impact (due to localized ranges, see Eken et al. 2004) and represent the single best indicators of areas of endemism and geographically unique allopatric speciation processes (see discussions in Raxworthy et al. 2007).

The Cerrado is a global biodiversity "hotspot" as defined by species richness, endemism, and human threats (Myers et al. 2000; Mittermeier et al. 2005). The region is being destroyed at a high rate, with 55% of its original vegetation already removed (Machado et al. 2004; Klink and Machado 2005). Given the urgency to conserve habitats and species, time to conduct adequate surveys of the entire region is not available. In such a scenario, ENM may prove to be a useful tool in conservation planning. Our results indicate that the use of maps provided by ENM may help to estimate species diversity, even when a geographical bias exists in the dataset used to generate the models. Still, we believe that ENM can be a useful tool to provide a big picture to guide survey efforts but may not be sufficient to justify management decisions and the fine-scale design of protected area systems. As in most of the Neotropical region, conservation opportunities lie in remote and generally poorly sampled regions. Data generated by ENM can maximize the utility of existing biodiversity data, providing a very useful first evaluation. However, for reliable conservation decisions ENM data must be followed by well-designed field inventories, especially for the detection of restricted range, rare species.

Acknowledgements We thank D. Shepard and A. T. Peterson for comments on the manuscript. Fieldwork was funded by Conservation International—Brazil, and field support was provided by Pequi, a Brazilian nongovernmental organization. Work on PNCM was authorized by IBAMA permit # 13204-1. We thank, P. Valdujo, S. Balbino, R. Recoder, and R. Bosque for help during fieldwork. This study was submitted in partial fulfilment of GCC's PhD degree at the University of Oklahoma. The species locality database was assembled as part of doctoral studies conducted by CN, supported by a FAPESP fellowship (# 02/00015-3). We thank J. Caldwell, M. Kaspari, J. Kelly, T. Rashed, and L. Vitt, for serving on GCC's doctoral committee and providing comments. GCC is supported by a Fulbright/CAPES PhD fellowship (15053155-2018/04-7). GRC by CNPq grant (# 302343/88-1). Portions of this research were supported by a National Science Foundation grant to Laurie J. Vitt and Janalee P. Caldwell (DEB-0415430).

References

Anderson RP, Peterson AT, Gómez-Laverde M (2002) Using niche-based GIS modeling to test geographic predictions of competitive exclusion and competitive release in South American pocket mice. Oikos 98:3–16

Anderson RP, Lew D, Peterson AT (2003) Evaluating predictive models of species' distributions: criteria for selecting optimal models. Ecol Model 162:211–232

Araújo MB, Luoto M (2007) The importance of biotic interactions for modelling species distributions under climate change. Global Ecol Biogeogr 16:743–753

Araújo MB, Rahbek C (2006) How does climate change affect biodiversity? Science 313:1396–1397

Austin MP (2007) Species distribution models and ecological theory: a critical assessment and some possible new approaches. Ecol Model 200:1–19

Brooks TM, da Fonseca GAB, Rodrigues ASL (2004) Species, data, and conservation planning. Conserv Biol 18:1682–1688

Chao A, Hwang WH, Chen YC et al (2000) Estimating the number of shared species in two communities. Stat Sin 10:227–246
Chazdon RL, Colwell RK, Denslow JS et al (1998) Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of NE Costa Rica. In: Dallmeier F, Comiskey JA (eds) Forest biodiversity research, monitoring, modeling: conceptual background, old world case studies. Parthenon Publishing, Paris, pp 285–309
Colli GR, Bastos RP, Araújo AFB (2002) The character and dynamics of the Cerrado herpetofauna. In: Oliveira PS, Marquis RJ (eds) The Cerrados of Brazil: ecology, natural history of a neotropical Savanna. Columbia University Press, New York
Colli GR, Giugliano LG, Mesquita DO et al (2009) A new species of *Cnemidophorus* from the Jalapão region, in the central Brazilian Cerrado. Herpetologica 65:311–327
Colwell RK (2005) EstimateS: statistical estimation of species richness and shared species from samples. User's guide and application published at http://purl.oclc.org/estimates
Colwell RK, Mao CX, Chang J (2004) Interpolating, extrapolating, and comparing incidence-based species accumulation curves. Ecology 85:2717–2727
Costa GC, Nogueira CC, Machado RB et al (2007) Squamate richness in the Brazilian Cerrado and its environmental–climatic associations. Divers Distrib 13:714–724
Costa GC, Wolfe C, Shepard DB et al (2008) Detecting the influence of climatic variables on species' distributions: a test using GIS niche-based models along a steep longitudinal environmental gradient. J Biogeogr 35:637–646
Díaz-Francés E, Soberón J (2005) Statistical estimation and model selection of species-accumulation functions. Conserv Biol 19:569–573
Domínguez-Domínguez O, Martinez-Meyer E, Zambrano L et al (2006) Using ecological niche modeling as a conservation tool for freshwater species: live-bearing fishes in central Mexico. Conserv Biol 20: 1730–1739
Duellman WE (1978) The biology of an equatorial herpetofauna in Amazonian Ecuador. Miscellaneous Publications of the University of Kansas, Museum of Natural History, Lawrence, pp 1–352
Eken G, Bennun L, Brooks TM et al (2004) Key biodiversity areas as site conservation targets. Bioscience 54:1110–1118
Elith J, Graham CH, Anderson RP et al (2006) Novel methods improve prediction of species' distributions from occurrence data. Ecography 29:129–151
Fielding AH, Bell JF (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ Conserv 24:38–49
Filho GAP, Montigelli GG (2006) Geographic distribution: *Hydrodynastes gigas*. Herpetol Rev 37:497
França FGR, Araújo AFB (2007) Are there co-occurence patterns that structure snake communities in central Brazil? Braz J Biol 67:33–40
Freitas MA, Silva TFS, Rodriguez MT (2007) Geographic distribution: *Chironius quadrilineatus*. Herpetol Rev 38:354
Garcia A (2006) Using ecological niche modelling to identify diversity hotspots for the herpetofauna of Pacific lowlands and adjacent interior valleys of Mexico. Biol Conserv 130:25–46
Guisan A, Broennimann O, Engler R et al (2006) Using niche-based models to improve the sampling of rare species. Conserv Biol 20:501–511
Guisan A, Graham CH, Elith J et al (2007a) Sensitivity of predictive species distribution models to change in grain size. Divers Distrib 13:332–340
Guisan A, Zimmermann NE, Elith J et al (2007b) What matters for predicting the occurrences of trees: techniques, data, or species' characteristics? Ecol Monogr 77:615–630
Heikkinen RK, Luoto M, Virkkala R et al (2007) Biotic interactions improve prediction of boreal bird distributions at macro-scales. Global Ecol Biogeogr 16:754–763
Hernandez PA, Graham CH, Master LL et al (2006) The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29:773–785
Hijmans RJ, Graham CH (2006) The ability of climate envelope models to predict the effect of climate change on species distributions. Global Change Biol 12:2272–2281
Hijmans RJ, Cameron SE, Parra JL et al (2005) Very high resolution interpolated climate surfaces for global land areas. Int J Climatol 25:1965–1978
Hortal J, Lobo JM, Jimenez-Valverde A (2007) Limitations of biodiversity databases: case study on seed-plant diversity in Tenerife, Canary Islands. Conserv Biol 21:853–863
Kadmon R, Farber O, Danin A (2004) Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. Ecol Appl 14:401–413
Klink CA, Machado RB (2005) Conservation of the Brazilian Cerrado. Conserv Biol 19:707–713

Lassueur T, Joost S, Randin CF (2006) Very high resolution digital elevation models: do they improve models of plant species distribution? Ecol Model 198:139–153
Lobo JM, Jiménez-Valverde A, Real R (2008) AUC: a misleading measure of the performance of predictive distribution models. Global Ecol Biogeogr 17:145–151
Loiselle BA, Jorgensen PM, Consiglio T et al (2008) Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? J Biogeogr 35:105–116
Lomolino MV (2004) Conservation biogeography. In: Lomolino MV, Heaney LR (eds) Frontiers of biogeography: new directions in the geography of nature. Sinauer, Sunderland, pp 293–296
Machado RB, Ramos Neto MB, Pereira PGP et al (2004) Estimativas de perda da área do Cerrado brasileiro. Conservation International Brasilia, Brasilia, DF
Mittermeier RA, Gil PR, Hoffman M et al (2005) Hotspots revisited: Earth's biologically richest and most endangered terrestrial ecoregions, 2nd edn. Conservation International, USA
Myers N (2003) Biodiversity hotspots revisited. Bioscience 53:916–917
Myers N, Mittermeier RA, Mittermeier CG et al (2000) Biodiversity hotspots for conservation priorities. Nature 403:853–858
Nogueira C (2001) New records of squamate reptiles in central Brazilian Cerrado II: Brasília region. Herpetol Rev 32:285–287
Nogueira C, Rodrigues MT (2006) The genus *Stenocercus* (Squamata: Tropiduridae) in extra-amazonian Brazil, with the description of two new species. South Am J Herpetol 1:149–165
Nogueira C, Colli GR, Martins M (2009) Local richness and distribution of the lizard fauna in natural habitat mosaics of the Brazilian Cerrado. Austral Ecol 34:83–96
Parra JL, Graham CC, Freile JF (2004) Evaluating alternative data sets for ecological niche models of birds in the Andes. Ecography 27:350–360
Pawar S, Koo MS, Kelley C et al (2007) Conservation assessment and prioritization of areas in Northeast India: priorities for amphibians and reptiles. Biol Conserv 136:346–361
Pearson RG, Raxworthy CJ, Nakamura M et al (2007) Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. J Biogeogr 34:102–117
Peterson AT (2006) Uses and requirements of ecological niche models and related distributional models. Biodivers Inform 3:59–72
Peterson AT, Nakazawa Y (2008) Environmental data sets matter in ecological niche modelling: an example with *Solenopsis invicta* and *Solenopsis richteri*. Global Ecol Biogeogr 17:135–144
Peterson AT, Vieglais DA (2001) Predicting species invasions using ecological niche modeling: new approaches from bioinformatics attack a pressing problem. Bioscience 51:363–371
Peterson AT, Papes M, Eaton M (2007) Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. Ecography 30:550–560
Peterson AT, Papes M, Soberon J (2008) Rethinking receiver operating characteristics analysis applications in ecological niche modeling. Ecol Model 213:63–72
Phillips SJ, Dudik M (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31:161–175
Phillips SJ, Dudik M, Shapire RE (2004) A maximum entropy approach to species distribution modeling. In: Greiner R, Schuurmans D (eds) Proceedings of the 21st international conference on machine learning. ACM Press Banff, Canada, pp 655–662
Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. Ecol Model 190:231–259
Poyry J, Luoto M, Heikkinen RK et al (2008) Species traits are associated with the quality of bioclimatic models. Global Ecol Biogeogr 17:403–414
R Development Core Team (2008) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org, ISBN 3-900051-07-0
Raes N, ter Steege H (2007) A null-model for significance testing of presence-only species distribution models. Ecography 30:727–736
Raxworthy CJ, Martinez-Meyer E, Horning N et al (2003) Predicting distributions of known and unknown reptile species in Madagascar. Nature 426:837–841
Raxworthy CJ, Ingram CM, Rabibisoa N et al (2007) Applications of ecological niche modeling for species delimitation: a review and empirical evaluation using day geckos (*Phelsuma*) from Madagascar. Syst Biol 56:907–923
Rodrigues MT, Pavan D, Curcio FF (2007) Two new species of lizards of the genus *Bachia* (Squamata, Gymnophthalmidae) from central Brazil. J Herpetol 41:545–553
Rodrigues MT, Camacho A, Nunes PMS et al (2008) A new species of the lizard genus *Bachia* (Squamata: Gymnophthalmidae) from the cerrados of central Brazil. Zootaxa 1875:39–50

Segurado P, Araújo MB (2004) An evaluation of methods for modelling species distributions. J Biogeogr 31:1555–1568

Silveira AL (2007) Geographic distribution: *Amphisbaena fuliginosa*. Herpetol Rev 38:481

Soberón J, Llorente J (1993) The use of species accumulation functions for the prediction of species richness. Conserv Biol 7:480–488

Soberón JM, Llorente JB, Onate L (2000) The use of specimen-label databases for conservation purposes: an example using Mexican Papilionid and Pierid butterflies. Biodivers Conserv 9:1441–1466

Soberón J, Jiménez R, Golubov J et al (2007) Assessing completeness of biodiversity databases at different spatial scales. Ecography 30:152–160

Stockwell DRB, Noble IR (1992) Induction of sets of rules from animal distribution data: a robust and informative method of data analysis. Math Comput Simul 33:385–390

Stockwell DRB, Peterson AT (2002) Effects of sample size on accuracy of species distribution models. Ecol Model 148:1–13

Strüssmann C, Carvalho MA (1998) New herpetological records for the state of Mato Grosso, western Brazil. Herpetol Rev 29:183–185

Sullivan BK (1981) Distribution and relative abundance of snakes along a transect in California. J Herpetol 15:245–246

Trivedi MR, Berry PM, Morecroft MD et al (2008) Spatial scale affects bioclimate model projections of climate change impacts on mountain plants. Global Change Biol 14:1089–1103

Valdujo PH, Nogueira CC, Baumgarten L et al (2009) Squamate reptiles from Parque Nacional das Emas and surroundings, Cerrado of Central Brazil. Check List 5:405–417

Whittaker RJ, Araújo MB, Paul J et al (2005) Conservation biogeography: assessment and prospect. Divers Distrib 11:3–23

Zweig MH, Campbell G (1993) Receiver-operating characteristics (ROC) plots—a fundamental evaluation tool in clinical medicine. Clin Chem 39:561–577