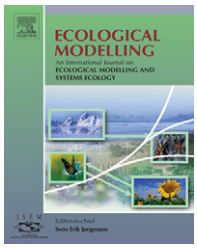
available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/ecolmodel



Rethinking receiver operating characteristic analysis applications in ecological niche modeling

A. Townsend Peterson, Monica Papeş, Jorge Soberón*

Natural History Museum and Biodiversity Research Center, The University of Kansas, Lawrence, KS 66045 USA

ARTICLE INFO

Article history:
Received 25 May 2007
Received in revised form
29 October 2007
Accepted 16 November 2007
Published on line 9 January 2008

Keywords:
Ecological niche model
Model evaluation
Receiver operating characteristic
Area under curve
Omission error

ABSTRACT

The area under the curve (AUC) of the receiver operating characteristic (ROC) has become a dominant tool in evaluating the accuracy of models predicting distributions of species. ROC has the advantage of being threshold-independent, and as such does not require decisions regarding thresholds of what constitutes a prediction of presence *versus* a prediction of absence. However, we show that, comparing two ROCs, using the AUC systematically undervalues models that do not provide predictions across the entire spectrum of proportional areas in the study area. Current ROC approaches in ecological niche modeling applications are also inappropriate because the two error components are weighted equally. We recommend a modification of ROC that remedies these problems, using partial-area ROC approaches to provide a firmer foundation for evaluation of predictions from ecological niche models. A worked example demonstrates that models that are evaluated favorably by traditional ROC AUCs are not necessarily the best when niche modeling considerations are incorporated into the design of the test.

The tools and techniques of ecological niche modeling (ENM) and the related ideas of species distribution modeling (SDM) have seen an impressive increase in activity in recent years (Guisan and Zimmermann, 2000; Soberón and Peterson, 2004; Araújo and Guisan, 2006). Many facets of these tools and their application have been examined in detailed analyses (Stockwell and Peterson, 2002a,b, 2003; Anderson et al., 2003; Pearson and Dawson, 2003; Araújo et al., 2005a,b; Guisan and Thuiller, 2005; Guisan et al., 2006; Pearson et al., 2007) that have greatly clarified the conditions of their use. However, in spite of such attention, the issue of how to evaluate predictions of these models statistically remains an area that is incompletely and unsatisfactorily resolved (Fielding and Bell, 1997; Araújo and Guisan, 2006; Guisan et al., 2006; Lobo et al., 2007).

In recent publications, statistical evaluations of niche and distribution model predictions have generally been based on receiver operating characteristic (ROC) analyses (DeLong et al., 1988), as exemplified by a recent, large-scale model comparison (Elith et al., 2006) and many similar studies. Spatial predictions can present errors of omission (false negatives, leaving out known distributional area) and errors of commission (false positives, including unsuitable areas in the prediction). ROC analysis involves plotting sensitivity (i.e., proportion of known presences predicted present, = 1 – false negative rate) against 1 – specificity (i.e., proportion of known absences predicted present, = false positive rate; Fig. 1). The area under the ROC curve (AUC) is then compared against null expectations [the area under the line linking the origin with upper right corner of the graph (1,1), = 0.5] either probabilistically or via bootstrap manipulations.

Here, we point out two sources of problems in ROC analyses that consistently favor certain kinds of algorithms over others. The first limitation of ROCs derives from the fact that certain algorithms span broad spectra of possible commission errors,

* *Corresponding author*. Tel.: +1 785 864 3926.
E-mail address: town@ku.edu (A.T. Peterson).
0304-3800/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2007.11.008

whereas others are restricted to smaller ranges—we show that ROCs consistently favor the former over the latter. The second limitation derives from the very different meanings of "absence" in the context of ENM *versus* SDM; as currently used, ROC analyses do not distinguish between the two, and, again, consistently favor model predictions oriented toward one type of analysis (SDM) over the other (ENM). We present a modification of the traditional ROC approach that takes steps towards resolving these two problems.

1. The (simple part of the) problem: unequal span of model predictions

A diverse set of inferential tools has been applied to the challenge of estimating niches and predicting geographic distributions of species (Elith et al., 2006; Peterson, 2006), ranging from simple range rules to complex neural networks, genetic algorithms, maximum entropy, and multivariate regression algorithms. The outputs from these different techniques have different characteristics: most relevant here is that different techniques may span very different ranges of predicted area of presence of a species (e.g., range rules predict one or a few thresholds, whereas multivariate regression approaches produce prediction across most of the spectrum of probabilities from 0 to 1). These differences, however, have implications for how AUC scores are calculated, because AUC calculations assume that 1 – specificity spans the entire range [0,1], even though model predictions may not span that whole range. Special modifications to the approach are required for development of AUC comparisons in partial ROCs that span only a subset of the full spectrum of areal predictions (Jiang et al., 1996; Dodd and Pepe, 2003).

ROC can be applied directly to evaluation of SDM predictions (Fielding and Bell, 1997; Fawcett, 2003; Phillips et al., 2006), although even this functionality is not above question (Lobo et al., 2007). A SDM produces a prediction value related (sometimes equal) to the probability that a species is present in a cell. By assigning thresholds, the continuous scores can be turned into binary predictions, which can be correct or incorrect, producing a contingency table called the "confusion matrix" (see Table 1). One confusion matrix exists per threshold value, and the four elements of the matrix can be used to calculate error characteristics.

In a conventional ROC, the proportion of true positives [$a/(a+c)$], equivalent to the sensitivity (or absence of omission error), is plotted against the proportion of false positives [$b/(b+d)$], which in turn is equivalent to 1 – specificity or the commission error. The plot in ROC space of sensitivity *versus* 1 – specificity displays how well an algorithm classifies instances as the threshold changes. In SDM and ENM applications, threshold changes mean that the area predicted as present also changes. Important sectors of this ROC space are the origin (0,0), where the algorithm never falsely identifies absences, but it fails to identify every known presence (which is useless); the top right corner (1,1), where the algorithm identifies every true presence correctly, but misidentifies all absences as positives (also useless, although in a different way). Finally, in the top left corner (0,1), the algorithm correctly identifies all true positives and never misclassifies a true absence as a presence. Therefore, the regions in ROC space near the (0,1) corner represent model predictions that successfully identify true presences and seldom misidentify absences as presences.

Now consider the behavior of a random classifier. Such an algorithm always randomly identifies as present a fixed proportion p of any set of instances, a function of the proportional area predicted present. This prediction rate is represented by the straight line joining the points (0,0) and (1,1). A random classificatory algorithm will select as present only a fraction p of true presences, giving a value of p on the sensitivity axis (y-axis). It will also select (wrongly) a fraction p of absences as presences, giving the same value of p on the x-axis. Therefore, as p varies, a line in which true presences = false presences is traced (Fig. 1).

The above ideas can be applied directly to situations in which true presences and true absences are known, such as the typical SDM problem (Guisan and Zimmermann, 2000). By varying the threshold at which the score of an algorithm is regarded as a presence, a curve in ROC space is traced (Fig. 1); elevation of this curve above the straight line of random expectation is a measure of the discrimination capacity of the algorithm (i.e., its capacity to classify correctly true presences and true absences) (Fielding and Bell, 1997; Guisan and Zimmermann, 2000). In an ENM context, however, the situation is slightly different, but different in important ways (see below).

In comparing the performance of different algorithms, in either a SDM or ENM context, a problem exists that – to our knowledge – has not been discussed previously in the literature on ENM or SDM: that some algorithms span the entire range of possible commission errors, while others cover only comparatively small regions of the overall ROC plot, either by design or by the intrinsic operation of the algorithm. In other words, while one algorithm may predict responses from 0 to 100% of false positives, another may predict only in the range of, for example, 40–90% (illustrated in Fig. 2 for Maxent, which predicts across the whole spectrum of areas, compared with GARP, which predicts only at the broader end of the spectrum, i.e., above ~60%; details of methodologies for model generation are provided below in the worked example). Note that the x-axis differs from that of a conventional ROC curve, an issue that will be discussed in detail below.

In practice, the ROC AUC is calculated based on a series of trapezoids (Fawcett, 2003), with the curve in essence "connecting the dots" in representing the different thresholds of

Table 1 – Schema of a confusion matrix, in which predicted presences and absences are related to their known status as observed presence or absence

	Observed	
	Present	Absent
Predicted		
Present	a	b
Absent	c	d

See text for explanation.

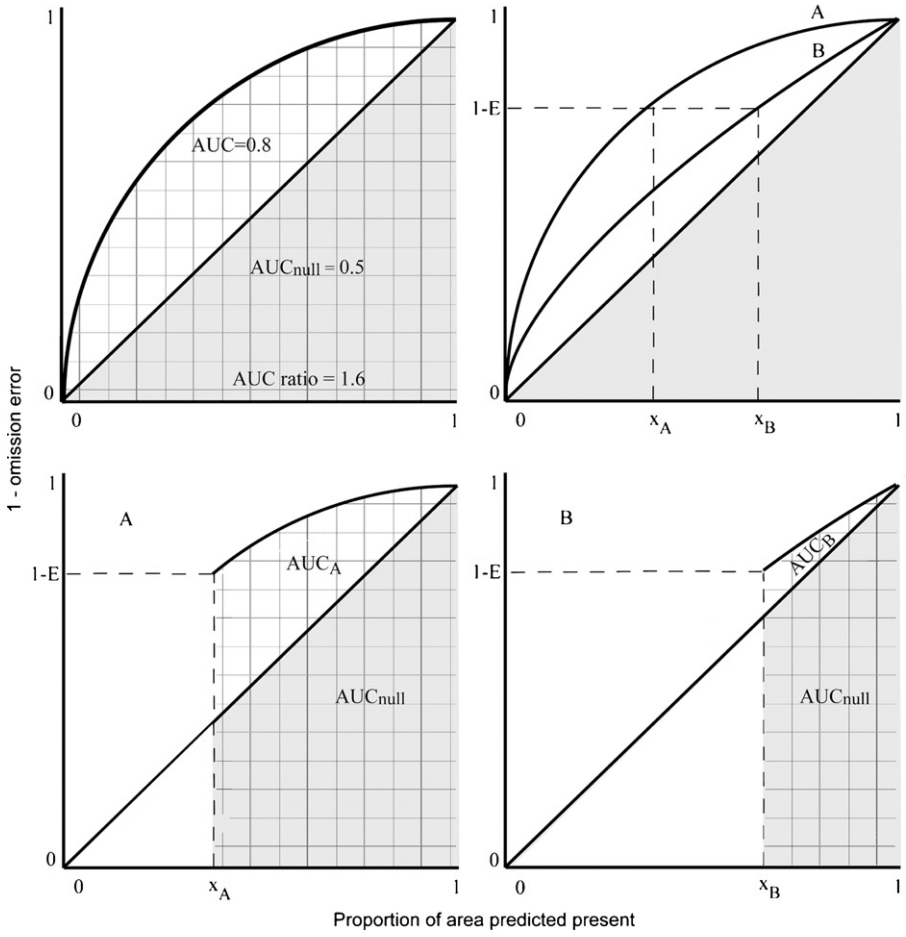


Fig. 1 – Summary of new recommendations for receiver operating characteristic (ROC) analysis in niche modeling. Upper left hand panel: traditional ROC approach, comparing the AUC of the test curve with 0.5, which is the AUC of the null expectation curve. Upper right hand panel: comparison of two curves, and illustration of how the user-chosen error tolerance *E* identifies different critical area thresholds for the two curves. Lower panels: illustration of the AUC comparisons that would be used to characterize each of the two curves.

the prediction. In the example in Fig. 2, Maxent has an AUC of 0.72 (ratio of observed to null expectations = 1.44), but GARP an AUC of only 0.63 (ratio = 1.26; Table 1)—the difference obtained because the first point of the GARP curve is automatically connected by a straight line to the origin. This procedure in effect penalizes algorithms with ROC curves that do not begin at or near the origin. In other words, in the example in Fig. 2, since GARP only predicts relatively broad geographic areas that have high rates of prediction of true presences (= low omission error) and does not make predictions at lower thresholds that would have higher omission errors, its ROC curve is defined only within a subset of possible areas. In this sense, we can now distinguish between two types of poor performance in ROC analyses: ROC curves that are genuinely lower and closer to the line of no information (AUC = 0.5), *versus* those that have artificially low AUC scores because they do not predict across the whole spectrum of proportional area predicted present. These complications are far from limited to GARP, however—BIOCLIM and related algorithms offer only a few thresholds of prediction, and many regression-based approaches have limited ranges of probabilities predicted.

2. Niche modeling considerations

The above is an artifactual problem that can affect any ROC analysis applied to analyses of different extents in predictions of proportional areas (Jiang et al., 1996; Dodd and Pepe, 2003). Another more subtle problem affects ROC analyses in ENM applications. Previous contributions have discussed differences between models of species' distributions and models of species' ecological niches (Soberón and Peterson, 2004, 2005; Peterson, 2006). Although seemingly a minor distinction, these differences have important implications for how model predictions should be evaluated.

Models of ecological niches are designed explicitly to predict *potential* areas of distribution, and therefore are generally broader than actual distributional areas (Hirzel et al., 2002; Soberón and Peterson, 2005; Phillips et al., 2006). Often, ENM applications are based on presence information only, owing quite simply to the practical lack of absence information, but even if absence data were to be available, they would have to be data regarding *absence from the potential distributional area*. As a consequence, data on absences of species

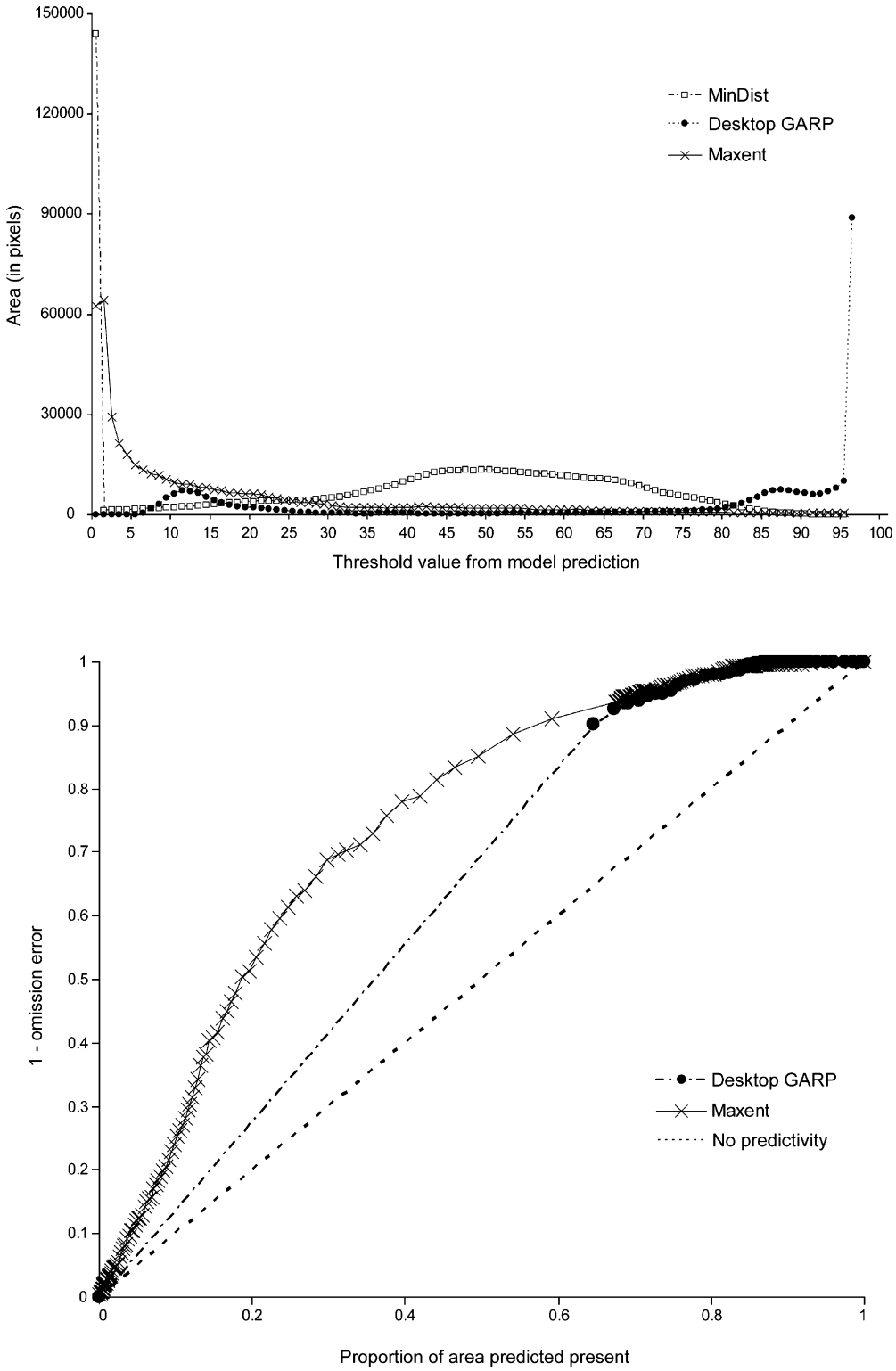


Fig. 2 – Summary of characteristics of three example model predictions. Top panel: area predicted present across the spectrum of thresholds for each model, based on a 3-threshold moving window. The bottom panel approximates closely a traditional receiver operating characteristic model, except that the *x*-axis is measured as proportion of the study area predicted present instead of being measured as success in predicting absence points; for simplicity, only GARP and Maxent results are shown in this panel.

are of dubious utility in the process of modeling ecological niches, unless some approach for generating more realistic absence data is used (Lobo et al., 2006). This point is easily understandable if one considers invasive species: if one had been modeling the species' ecological niche a few decades prior to its introduction in the novel region, one would have counted its future adventive distributional area as absence. This area was actually quite within the ecological niche dimensions of the species, as was demonstrated by the later invasion.

As such, unless obtained somehow from the potential distribution (e.g., if annual plants were seeded experimentally across the region), absence data should not be employed in evaluating model quality in ENM applications. For this reason, and following previous observations by Phillips et al. (Phillips et al., 2006) that the logic of ROC allows for more general partitioning of instances than "presences" versus "absences," we follow a modified ROC procedure that disposes entirely of absence data. Rather, we calculate the values used as the *x*-axis as the proportion of the overall area predicted as present, rather than using commission error calculated based on vaguely defined (and often unavailable) data summarizing "absences" (Phillips et al., 2006).

Since absence data have been omitted and the 1 – specificity axis changed to proportion of area predicted as present, new interpretations are needed. In previous analyses, in which niche models were evaluated via expert opinion (Anderson et al., 2003), it was shown that omission error characteristics are more important in distinguishing good from bad models than are commission error considerations. Put simply, in a niche-modeling framework, a model that errs by omitting known points of presence is more seriously flawed than one that predicts areas not known to be inhabited (Raxworthy et al., 2003). Among models that overpredict, what is more, some 'overprediction' is in disjunct areas that likely represent areas inaccessible to the species for reasons unrelated to landscape suitability (e.g., historical dispersal limitations, speciation events, interspecific interactions) (Peterson et al., 1999; Wiens, 2004; Wiens and Graham, 2005): these areas do not represent model prediction error, but rather offer an accurate depiction of the spatial extent of habitable conditions for the species. Other models may reconstruct overly broad suites of environmental conditions as suitable for the species—these models genuinely fail in reconstructing the ecological niche of the species because they do not distinguish effectively between potential presence and absence. Distinguishing between these two possibilities (predicting areas not inhabited for nonecological reasons *versus* predicting an overly broad suite of environmental conditions) represents an important ongoing priority in the development of this field, and depends in large part on being able to decide which models are "better" than others.

3. Modified ROC approach

Given the above considerations, we outline a series of modifications to ROC analysis that make it consistent with the characteristics of ENM applications, building on previous work with partial-area ROC analyses, as follows. (1) The *x*-axis is not calibrated based on successful *versus* unsuccessful prediction of absence points, but rather on the proportion of the overall area under consideration predicted as present. This change follows the reasoning that Phillips et al. (Phillips et al., 2006) used in substituting "background points" for "absence points" in their analyses of the Maxent algorithm. (2) AUC calculations are restricted to the domain of prediction of the algorithm, and do not extend to intervals along the *x*-axis in which an algorithm does not make predictions. Finally, (3) we restrict AUC calculations to the domain within which omission error is sufficiently low as to meet user-defined requirements of predictive ability (Pepe, 2000). In this section, we develop these latter ideas in detail, and then illustrate the differences in a worked example.

We begin by defining a user-selected parameter E, which refers to the amount of error admissible along the true-positives axis, given the requirements and conditions of the study. This parameter refers to how much omission error is acceptable—it might be set at $E = 0$ in applications in which highest-quality occurrence data are used, or it might be higher (perhaps 5–20%) when the occurrence data are known to include certain amounts of error (e.g., when using "found" data). Hence, the researcher considers the error characteristics of the data that will be used to test the model predictions and the needs of the particular study, and chooses a value of E appropriate to the question at hand.

Fig. 1 illustrates these ideas graphically: the upper left-hand panel depicts a typical ROC analysis, in which a curve representing some model prediction has an AUC = 0.8, which is then compared to the AUC for a line of null expectations (= 0.5) and significance values are obtained either by combinatorial probability calculations or by bootstrapping (DeLong et al., 1988; Vida, 2006). The upper right-hand panel shows two such curves, one of which (curve A) is clearly 'better' than the other (curve B), in that it is more elevated from the line of null expectations.

In our proposed modification, the line defined by $1 - E$ on the vertical axis is intersected with the two ROC curves, and the projection of each to the *x*-axis is used to identify key area thresholds for the models, in this case x_A and x_B (Fig. 1). The lower 2 panels of Fig. 1 show the AUC comparisons for the ROC curves that would be used in our modified comparisons. In each, we consider only the portion of the ROC curve that lies within the predictive range of the modeling algorithm and within the range of acceptable models in terms of omission error ($1 - E$ to 1). Also in each, the null expectations of AUC are <0.5 because only part of the full range of proportional areas predicted present is included in the calculations. The area under the ROC curve for each model can then be calculated empirically as a series of trapezoids (DeLong et al., 1988; Burden and Faires, 2005). Given both the change in the definition of the *x*-axis and the now-variable AUC for the null expectation, we now express ROC results as ratios of the area under the observed curve to the area under the trapezoid defined by the random line and the interval x_A (or x_B) to 1. This value departs from unity as the model's ROC curve improves with respect to random expectations, and comparisons of model ROC AUCs with null expectations must be achieved by means of bootstrapping.

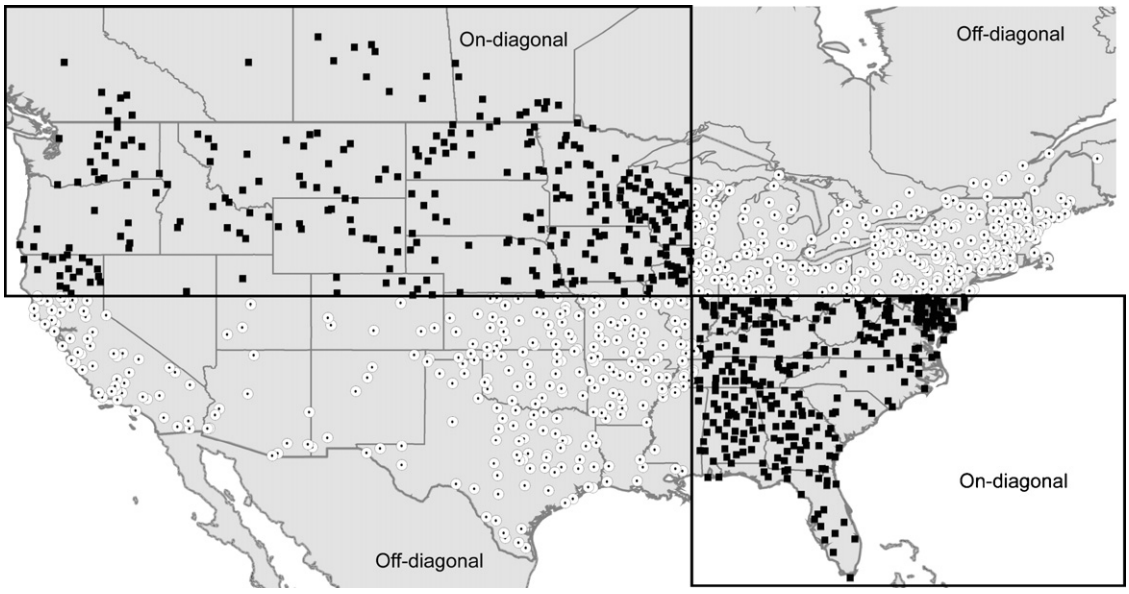


Fig. 3 – Occurrence data used in the example discussed in the text: black and white points are occurrences of Mourning Dove (*Zenaida macroura*) drawn from the North American Breeding Bird Survey (1991–2000). Models were built based on the points in the off-diagonal quadrants, and were tested based on the points in the on-diagonal quadrants.

4. Worked example

4.1. *Methods*

We use here an example analysis drawn from a recent comparative study (Peterson et al., 2007). As full details are provided in that publication, and given that the points made herein are not specific to any particular methodology, we here only provide a sketch of the methods that were employed. We based this example on Mourning Dove (*Zenaida macroura*) occurrence data drawn from the North American Breeding Bird Survey (BBS) (Sauer et al., 2001). To assure that occurrences used in analyses represent reasonably stable populations, we used only BBS survey routes on which the species had been detected in ≥8 years in 1991–2000; overall, 1202 presence points were available for the analyses.

To challenge the ENM algorithms to predict into broad unsampled areas (a niche-modeling challenge), we separated available occurrence points into quadrants based on whether their coordinates fell above or below the median longitude and median latitude of occurrence localities. Henceforth, we refer to the NW and SE quadrants as 'on-diagonal,' and the NE and SW pair of quadrants as 'off-diagonal' (Fig. 3); we trained models based on off-diagonal quadrants (582 points) and tested them using the independent occurrence points in the on-diagonal quadrants (620 points) (Peterson and Shaw, 2003); this manipulation challenges modeling algorithms to predict into unsampled regions, rather than simply interpolating or filling gaps in a densely sampled landscape. It is important to note that all aspects of model development (including, e.g., best subsets filtering in GARP) (Anderson et al., 2003) were carried out on one pair of quadrants, and testing and model evaluation in the other pair of quadrants only.

We characterized North American (24.3–76.5°N, 52.0–169.5°W) environments based on 19 biologically meaningful climate parameters drawn from the 10′ WorldClim data set (Hijmans et al., 2005), supplemented with information on topographic features summarized in four additional raster data layers (elevation, slope, aspect, compound topographic index) from the 1 km resolution Hydro-1K digital elevation model data set (USGS, 2001). All data sets were resampled to 10′ resolution to reflect the spatial accuracy of the occurrence data; the dimensionality of the environmental data was reduced by means of principal components analysis (PCA) to create new axes that summarized variation in fewer (independent) dimensions (Peterson et al., 2007). We retained the first 11 components, which together explained >99% of the overall variation in environmental parameters.

Several approaches have been used to approximate species' ecological niches (Segurado and Araújo, 2004), as exemplified by a recent broad comparative study of model performance (Elith et al., 2006). Here, for the purpose of illustration, we compared three methods: one that performed relatively poorly in the Elith et al. (2006) study, the Genetic Algorithm for Rule-set Prediction (GARP) (Stockwell, 1999; Pereira, 2002), *versus* one of the top performers, a maximum entropy (Maxent) approach (Phillips et al., 2006). Also included was the Minimum Distance algorithm (OpenModeller[1], version 0.1; hereafter MinDist), which is equivalent to the simplest manifestation of DOMAIN, but based on Euclidean distances instead of on the Gower metric (Carpenter et al., 1993), as this method presents some interesting contrasts with the other two algorithms.

GARP models were developed using a desktop version that permits flexibility in model development (Pereira, 2002).

[1] http://openmodeller.sourceforge.net/.

In GARP, occurrence points from the pair of quadrants on which models are to be trained are divided randomly into training and "extrinsic test data" sets; the former is again divided evenly into "training data" (for model rule development) and "intrinsic test data" sets (for model rule evaluation and refinement). GARP works in an iterative process of rule selection, evaluation, testing, and incorporation or rejection: first, a method is chosen from a set of possibilities, and then is applied to the training data and a rule developed; rules may evolve by a number of means (e.g., truncation, point changes, crossing-over among rules) to maximize predictivity. Predictive accuracy is then evaluated independent points resampled from the intrinsic test data, and change in predictive accuracy from one iteration to the next is used to evaluate whether a particular rule should be incorporated into the model. To force GARP models to be general, and to minimize overfitting, following procedures in all recent GARP applications, we used the best subsets procedure (Anderson et al., 2003). We then summed the resulting 100 grids to create a surface summarizing model agreement, with values ranging 0–100 as integers.

Maxent models were developed using software described and tested in detail in recent publications (Phillips et al., 2004, 2006). Maxent focuses on fitting a probability distribution for occurrence of the species in question to the set of pixels across the study region, based on the idea that the best explanation for unknown phenomena will maximize the entropy of the probability distribution, subject to the appropriate constraints. In the case of modeling ecological niches of species, these constraints consist of maintaining the difference between the mean values of the variable distributions predicted by the algorithm and the observed means always smaller than a "regularization parameter," β (Phillips et al., 2004, 2006). We used default parameters for Maxent models (i.e., no random subsampling, regularization multiplier = 1500 maximum iterations, 10,000 background points, convergence limit = 10^{-5}). Given the real-number nature of Maxent predictions, and given the much-greater ease of manipulation of integer grids, we imported results into ArcView as floating-point grids, multiplied them by 100, and converted them to integer grids for further analysis.

Finally, we estimated niche models using the very simple MinDist algorithm. Here, for each pixel in the landscape, the Euclidean distance in a normalized environmental space is calculated to each known occurrence point. The minimum of this set of distance measures is assigned as the predicted value of the pixel in question. Although a maximum distance parameter can be set to eliminate very large distances from consideration, we did not make any such assumptions, and rather allowed each pixel in the landscape to be assigned a continuous variable that indicates similarity to known occurrences of the species.

We summarized these three models in various manners that relate to ROC analyses, all based solely on independent testing points from the quadrants that were not used to train the models. In particular, at each predictive threshold, we calculated sensitivity as 1 − omission error, the latter measured based on the independent testing data from the other two quadrants of the species' distribution (Fig. 3). We calculated AUCs using the trapezoid method (Burden and Faires, 2005), and present our AUC comparisons as the ratio of the

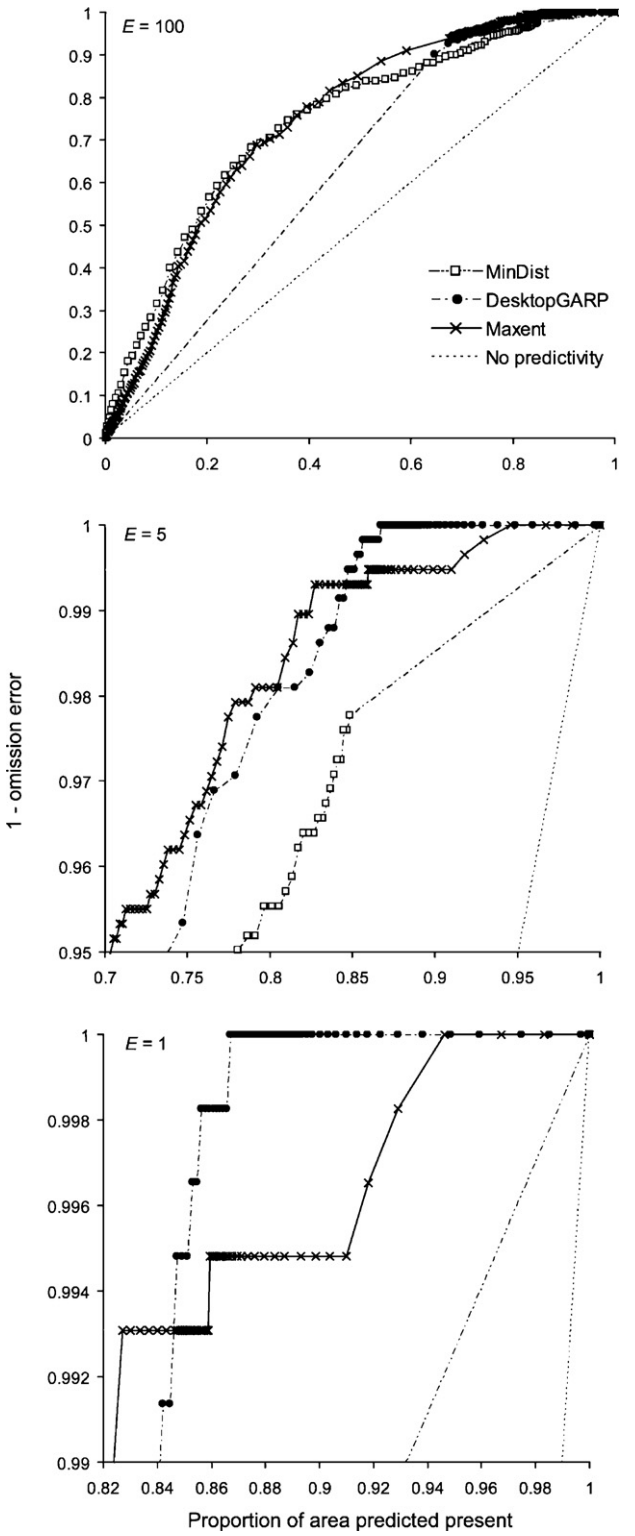


Fig. 4 – Illustration of receiver operating characteristic (ROC) curves at different thresholds of *E*, the user-defined error tolerance. *E* = 100 is equivalent to the traditional ROC analysis, but the lower two panels show *E* = 5 and 1. The point is that the relationships between the curves change as one focuses on the lower-omission models instead of the whole spectrum of thresholds.

Table 2 – Summary of statistics describing receiver operating characteristic curves for three modeling algorithms (MinDist, GARP, Maxent) at each of three values of *E*, the threshold of acceptable omission error (MinDist statistics only presented for two *E* values, for reasons explained in the text)

	MinDist		GARP			Maxent		
	$E = 100$	$E = 5$	$E = 100$	$E = 5$	$E = 1$	$E = 100$	$E = 5$	$E = 1$
Minimum	1.406	1.048	1.248	1.119	1.080	1.439	0.982	0.953
Maximum	1.500	1.130	1.304	1.183	1.122	1.539	1.179	1.118
Mean	1.457	1.093	1.273	1.146	1.087	1.488	1.132	1.060
Standard deviation	0.020	0.018	0.010	0.014	0.007	0.018	0.049	0.041
Number of replicates ≤ 1	0	0	0	0	0	0	18	25
P	9.8×10^{-114}	2.7×10^{-7}	1.0×10^{-174}	8.1×10^{-25}	8.2×10^{-34}	1.3×10^{-156}	0.0032	0.0736

Values presented are AUC ratios (minimum, maximum, mean, and standard deviation) across 200 bootstrap replicates, the number of bootstrap replicates falling at or below unity, and the probability that the mean is ≤ 1 based on a standard normal variate associated with the mean and standard deviation.

model AUC to the null expectation described above (referred to henceforth as "AUC ratios"). Bootstrapping manipulations to permit evaluation of statistical significance of AUCs (as compared with null expectations) were achieved by resampling 290 test points (50% of the total test points available) with replacement 200 times from the overall pool of testing data in S-Plus (version 7); one-tailed significance of differences in AUC from the line of null expectations was assessed both via fitting a standard normal variate (the z-statistic) and calculating the probability that the mean AUC ratio is ≤ 1, and separately by counting the number of bootstrap replicates with AUC ratios of ≤ 1.

4.2. *Application*

We developed modified ROC curves for each of the three model outputs at each of three values of *E*: 100% (in which the user accepts models across the entire spectrum of areas predicted as present, equivalent to the traditional ROC application), 5, and 1% (Fig. 4, Table 2). As mentioned above, at $E = 100\%$, Maxent clearly outperformed GARP, as did MinDist. However, it is clear in Fig. 4 that much of this difference springs from the fact that the GARP model makes no predictions of less than ~65% of the study area: the chord drawn from that point on the graph to the origin leaves out much of the area included under the Maxent and MinDist curves. AUC ratios were 1.46 for MinDist and 1.49 for Maxent, but only 1.27 for GARP, although all three were significantly elevated above the line of null expectations (bootstrap manipulation, all $P \ll 0.05$; Table 1).

At $E = 5\%$, however, the relative positions of the curves shift. Ignoring the lower part of the curves (corresponding to model thresholds that omit more than the user-stated tolerance), now Maxent and GARP are the higher curves, and MinDist is considerably lower (Fig. 4, Table 2). AUC ratios were highest for Maxent and GARP (1.13 and 1.15, respectively), and lower for MinDist (1.09). Although all three were statistically significantly better than null expectations using the z-statistic, Maxent did not achieve statistical significance based on the simpler counts of numbers of replicates with AUC ratios of > 1 (Table 2). It is worthy of note that MinDist provides predictions only up to 84.8% of the study area, and so the region between that value and unity along the *x*-axis was filled by a straight line; the incomplete trace of the ROC curve provided by MinDist is unlikely to account for its poor performance in the partial-area AUC calculation, and we retain it in this example for full comparability with the other two methods. Full implementation of the methodology we present may wish to limit comparisons in this region of the graph as well.

Finally, at $E = 1\%$, MinDist provided no predictions in this interval (Fig. 4), and so was excluded from calculations. GARP had an AUC ratio of 1.09, as compared with Maxent's 1.06. While the GARP AUC was significantly higher than null expectations by both measures, the Maxent AUC ratio was ≤ 1 in 12% of the bootstrap replicates, and the z-statistic yielded a $P = 0.074$, indicating that the Maxent curve was not significantly elevated above the null expectations (Table 2). Hence, models that appeared most accurate in their predictions at $E = 100$ were not the most accurate when model tests were restricted to the region of interest in the niche modeling exercise.

5. Discussion and conclusions

We emphasize that the purpose of this contribution is not to establish that any niche modeling method is better or worse than any other method. In fact, we considered removing modeling method names from the manuscript and replacing them with "X," "Y," and "Z," to focus readers' attention on the key points. In particular, we assert that currently accepted model evaluation techniques are not adequate for niche modeling applications, and can yield inaccurate and inappropriate conclusions in many cases. This paper presents a first set of steps towards remedying these failings.

A recent broad comparison (in which two of the authors of this paper participated) compared 14 modeling methods, and identified a suite of methods with particularly good predictive abilities that included Maxent (Elith et al., 2006). This large-scale comparison was nonetheless designed as an SDM (distribution-modeling) exercise, and as such included absence data as an integral element in model testing. The three measures of model predictivity employed (the customary ROC analyses, a kappa statistic, and a correlation-based procedure) all balance correct predictions of presences and absences, measuring the ability of an algorithm to discriminate between sites where a species is present and those

where it is absent (Elith et al., 2006). However, as demonstrated above, traditional ROC approaches can identify a method as highly accurate when the method in fact is *inferior* in the range of predictive thresholds that is likely to be of interest in niche modeling exercises (this point indicates that *part* of the variation among models in the study in question is artifactual, regardless of whether the application is one of SDM or one of ENM). Furthermore, we point out that many recent applications of these methods are *explicitly* ENM (niche-modeling) applications (Anderson et al., 2002; Pearson et al., 2002; Graham et al., 2004; Araújo et al., 2005a; Thuiller et al., 2005; Wiens and Graham, 2005), so the customary ROC analyses should be regarded with caution in these cases.

The evaluation methodology outlined in this paper achieves several of our goals. First, it removes the emphasis on absence data, which in niche-modeling applications can be positively misleading (Peterson, 2006). Second, it emphasizes the key role of omission error in evaluating niche model predictivity (Anderson et al., 2003). Finally, we follow a previous suggestion in a very different application of ROC analysis that analyses may best be limited to subsectors of the ROC space when certain portions of that space are not directly relevant to applications of interest (Jiang et al., 1996; Pepe, 2000; Dodd and Pepe, 2003)—in niche modeling, this modification allows the user to set bounds on the types of predictions that are to be considered (Pepe, 2000; Dodd and Pepe, 2003). A researcher interested in evaluating the invasive potential of a species would almost certainly be disappointed with a method that performs well at thresholds that have associated omission errors of >50%! Taking the intended uses of the model into account, as well as the error-related characteristics of the input data, is an important refinement to model evaluation approaches.

Clearly, much work remains in the development of these methodologies. In reality, limits both to the sensitivity and the false-positive axes may be desirable (Jiang et al., 1996; Dodd and Pepe, 2003). In our modified approach, limiting the fraction of total area that an algorithm is allowed to predict (the overprediction) may be biologically sensible. This restriction would create partial ROC curves, limited on one side by the minimum sensitivity acceptable, and on the other by the maximum overprediction that is tolerable. When more experience in the use of our modified approach is gathered, development of a broad, comparative study parallel to the previous (distribution modeling) study (Elith et al., 2006), but based on niche modeling ideas, would be particularly instructive. Because the approach we present here is distinct in several ways from conventional ROC analysis, the probabilistic interpretations of ROC scores (Fawcett, 2003) and its relations with the Mann–Whitney test will need to be reassessed. Finally, once the final form of the methodology for partial-ROC applications to predictions of species' geographic distributions is clear, developing program code to permit easy implementation would be desirable.

Acknowledgements

We thank many valued colleagues for discussions of these and related topics over the past several years, particularly Enrique Martínez-Meyer, Robert Anderson, Richard Pearson, Jake Overton, and Simon Ferrier, although we may or may not have agreed.

REFERENCES

Anderson, R.P., Lew, D., Peterson, A.T., 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models. Ecol. Model. 162, 211–232.

Anderson, R.P., Peterson, A.T., 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 93, 3–16.

Araújo, M.B., Guisan, A., 2006. Five (or so) challenges for species distribution modelling. J. Biogeogr. 33, 1677–1688.

Araújo, M.B., Pearson, R.G., Thuiller, W., Erhard, M., 2005a. Validation of species-climate impact models under climate change. Global Change Biol. 11, 1504–1513.

Araújo, M.B., Whittaker, R.J., Ladle, R.J., Erhard, M., 2005b. Reducing uncertainty in projections of extinction risk from climate change. Global Ecol. Biogeogr. 14, 529–538.

Burden, R.L., Faires, J.D., 2005. Numerical Analysis, eighth ed. Thomson Books, Belmont, California.

Carpenter, G., Gillison, A.N., Winter, J., 1993. DOMAIN: a flexible modeling procedure for mapping potential distributions of animals and plants. Biodivers. Conserv. 2, 667–680.

DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L., 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44, 837–845.

Dodd, L.E., Pepe, M.S., 2003. Partial AUC estimation and regression. Biometrics 59, 614–623.

Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettman, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S.E., Wisz, M.S., Zimmermann, N.E., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29, 129–151.

Fawcett, R., 2003. ROC Graphs: Notes and Practical Considerations for Data Mining Research. Technical Report HPL-2003-4. HP Laboratories, Palo Alto, California.

Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 24, 38–49.

Graham, C.H., Ron, S.R., Santos, J.C., Schneider, C.J., Moritz, C., 2004. Integrating phylogenetics and environmental niche models to explore speciation mechanisms in dendrobatid frogs. Evolution 58, 1781–1793.

Guisan, A., Lehmann, A., Ferrier, S., Austin, M.P., Overton, J.M., Aspinall, R., Hastie, T., 2006. Making better biogeographical predictions of species' distributions. J. Appl. Ecol. 43, 386–392.

Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecol. Lett. 8, 993–1009.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Modell. 135, 147–186.

Hijmans, R.J., Cameron, S., Parra, J., 2005. WorldClim, Version 1.3, http://biogeo.berkeley.edu/worldclim/worldclim.htm. University of California, Berkeley.

Hirzel, A.H., Hausser, J., Chessel, D., Perrin, N., 2002. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? Ecology 83, 2027–2036.

Jiang, Y., Metz, C.E., Nishikawa, R.M., 1996. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. Radiology 201, 745–750.
Lobo, J.M., Jimenez-Valverde, A., Real, R., 2007. AUC: a misleading measure of the performance of predictive distribution models. Global Ecol. Biogeogr..
Lobo, J.M., Verdú, J.R., Numa, C., 2006. Environmental and geographical factors affecting the Iberian distribution of flightless *Jekelius* species (Coleoptera: Geotrupidae). Divers. Distributions 12, 179–188.
Pearson, R.G., Dawson, T.P., 2003. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? Global Ecol. Biogeogr. 12, 361–371.
Pearson, R.G., Dawson, T.P., Berry, P.M., Harrison, P.A., 2002. SPECIES: a spatial evaluation of climate impact on the envelope of species. Ecol. Modell. 154, 289–300.
Pearson, R.G., Raxworthy, C., Nakamura, M., Peterson, A.T., 2007. Predicting species' distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. J. Biogeogr. 34, 102–117.
Pepe, M.S., 2000. Receiver operating characteristic methodology. J. Am. Stat. Assoc. 95, 308–311.
Pereira, R.S., 2002. Desktop GARP. http://www.lifemapper.org/desktopgarp/.
Peterson, A.T., 2006. Uses and requirements of ecological niche models and related distributional models. Biodivers. Inform. 3, 59–72.
Peterson, A.T., Papeş, M., Eaton, M., 2007. Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. Ecography 30, 550–560.
Peterson, A.T., Shaw, J.J., 2003. *Lutzomyia* vectors for cutaneous leishmaniasis in southern Brazil: ecological niche models, predicted geographic distributions, and climate change effects. Int. J. Parasitol. 33, 919–931.
Peterson, A.T., Soberón, J., Sánchez-Cordero, V., 1999. Conservatism of ecological niches in evolutionary time. Science 285, 1265–1267.
Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecol. Modell. 190, 231–259.
Phillips, S.J., Dudik, M., Schapire, R.E., 2004. A maximum entropy approach to species distribution modeling. In: Proceedings of the 21st International Conference on Machine Learning.
Raxworthy, C.J., Martínez-Meyer, E., Horning, N., Nussbaum, R.A., Schneider, G.E., Ortega-Huerta, M.A., Peterson, A.T., 2003. Predicting distributions of known and unknown reptile species in Madagascar. Nature 426, 837–841.
Sauer, J.R., Hines, J.E., Fallon, J., 2001. The North American Breeding Bird Survey, Results and Analysis 1966–2000, version 2001.2. USGS Patuxent Wildlife Research Center, Laurel, MD.
Segurado, P., Araújo, M.B., 2004. An evaluation of methods for modelling species distributions. J. Biogeogr. 31, 1555–1568.
Soberón, J., Peterson, A.T., 2004. Biodiversity informatics: managing and applying primary biodiversity data. Philos. Trans. R. Soc. Lond. B 359, 689–698.
Soberón, J., Peterson, A.T., 2005. Interpretation of models of fundamental ecological niches and species' distributional areas. Biodivers. Inform. 2, 1–10.
Stockwell, D.R.B., 1999. Genetic algorithms II. In: Fielding, A.H. (Ed.), Machine Learning Methods for Ecological Applications. Kluwer Academic Publishers, Boston, pp. 123–144.
Stockwell, D.R.B., Peterson, A.T., 2002a. Controlling bias in biodiversity data. In: Scott, J.M., Heglund, P.J., Morrison, M.L. (Eds.), Predicting Species Occurrences: Issues of Scale and Accuracy. Island Press, Washington, DC, pp. 537–546.
Stockwell, D.R.B., Peterson, A.T., 2002b. Effects of sample size on accuracy of species distribution models. Ecol. Modell. 148, 1–13.
Stockwell, D.R.B., Peterson, A.T., 2003. Comparison of resolution of methods used in mapping biodiversity patterns from point occurrence data. Ecol. Indicators 3, 213–221.
Thuiller, W., Richardson, D.M., Pysek, P., Midgley, G.F., Hughes, G.O., Rouget, M., 2005. Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. Global Change Biol. 11, 2234–2250.
USGS, 2001. HYDRO1k Elevation Derivative Database, http://edcdaac.usgs.gov/gtopo30/hydro/. U.S. Geological Survey, Washington, D.C.
Vida, S., 2006. Accumetric Test Performance Analysis, Version 1.1. Accumetric Corporation, Montreal.
Wiens, J.J., 2004. Speciation and ecology revisited: phylogenetic niche conservatism and the origin of species. Evolution 58, 193–197.
Wiens, J.J., Graham, C.H., 2005. Niche conservatism: integrating evolution, ecology, and conservation biology. Annu. Rev. Ecol. Evol. Syst. 36, 519–539.