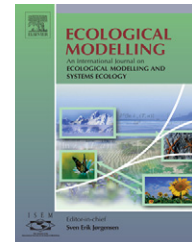
ECOLOGICAL MODELLING 213 (2008) 63–72



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.