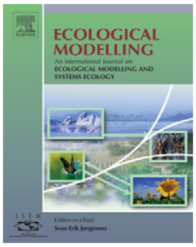
available at www.sciencedirect.com

ScienceDirect

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



A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa

Elizabeth A. Freeman[*], *Gretchen G. Moisen*

USDA Forest Service, Rocky Mountain Research Station, 507 25th Street, Ogden, UT 84401, USA

ARTICLE INFO

Article history:
Received 24 July 2007
Received in revised form
30 April 2008
Accepted 14 May 2008
Published on line 4 July 2008

Keywords:
Binary classification
ROC
AUC
Sensitivity
Specificity
Threshold

ABSTRACT

Modelling techniques used in binary classification problems often result in a predicted probability surface, which is then translated into a presence–absence classification map. However, this translation requires a (possibly subjective) choice of threshold above which the variable of interest is predicted to be present. The selection of this threshold value can have dramatic effects on model accuracy as well as the predicted prevalence for the variable (the overall proportion of locations where the variable is predicted to be present). The traditional default is to simply use a threshold of 0.5 as the cut-off, but this does not necessarily preserve the observed prevalence or result in the highest prediction accuracy, especially for data sets with very high or very low observed prevalence. Alternatively, the thresholds can be chosen to optimize map accuracy, as judged by various criteria. Here we examine the effect of 11 of these potential criteria on predicted prevalence, prediction accuracy, and the resulting map output. Comparisons are made using output from presence–absence models developed for 13 tree species in the northern mountains of Utah. We found that species with poor model quality or low prevalence were most sensitive to the choice of threshold. For these species, a 0.5 cut-off was unreliable, sometimes resulting in substantially lower kappa and underestimated prevalence, with possible detrimental effects on a management decision. If a management objective requires a map to portray unbiased estimates of species prevalence, then the best results were obtained from thresholds deliberately chosen so that the predicted prevalence equaled the observed prevalence, followed closely by thresholds chosen to maximize kappa. These were also the two criteria with the highest mean kappa from our independent test data. For particular management applications the special cases of user specified required accuracy may be most appropriate. Ultimately, maps will typically have multiple and somewhat conflicting management applications. Therefore, providing users with a continuous probability surface may be the most versatile and powerful method, allowing threshold choice to be matched with each maps intended use.

Published by Elsevier B.V.

1. Introduction

Binary classification mapping is a technique crucial to multiple areas of study. Applications include mapping species distribution, disturbance, wildlife habitat, insect and disease outbreaks, fire risk, and climate change. Modelling techniques often generate predictions that are analogous to a probability of presence. A common practice is to translate this surface

[*] *Corresponding author*. Tel.: +1 801 510 3765.
E-mail address: eafreeman@fs.fed.us (E.A. Freeman).

0304-3800/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.ecolmodel.2008.05.015

into a simple 0/1 classification map by a choice of threshold, or cut-off probability, beyond which something is classified as present. The selection of this threshold value can have dramatic effects on model accuracy as well as the predicted prevalence (the overall proportion of locations where the variable is predicted to be present). The traditional default is to simply use a threshold of 0.5 as the cut-off, but this does not necessarily preserve the observed prevalence or result in the highest prediction accuracy, especially for data sets with very high or very low observed prevalence.

Alternatively, the thresholds can be chosen to optimize map accuracy, as judged by one of several criteria. Because the utility of maps for different management applications cannot be captured in a single map accuracy number, several global measures are commonly used to assess the predictive performance of the models; these include percent correctly classified (PCC), sensitivity, specificity, kappa, and receiver operating curves (ROC plots), with their associated area under the curve (AUC). In addition, in many applications, it is important that the predicted prevalence reflects the observed prevalence, and agreement between these two may also be used as a measure of map accuracy. All of these numerous accuracy measures have been used as in various ways to create criteria for threshold optimization, as described below.

Beginning with the simplest of these measures, PCC is the proportion of test observations that are correctly classified. However this can be deceptive when prevalence is very low or very high. For example, species with very low prevalence, it is possible to maximize PCC simply by declaring the species absent at all locations, resulting in a map with little usefulness. Although sometimes used to optimize threshold values, the accuracy measure itself has little value in practice.

As a result, classification accuracy is commonly broken down into two measures. Sensitivity, or proportion of correctly predicted positive observations, reflects a model's ability to detect a presence, given a species actually occurs at a location. Specificity, or proportion of correctly predicted negative observations, reflects a model's ability to predict an absence where a species does not exist. Sensitivity and specificity can be combined in various ways to assess model quality and optimize thresholds. Fielding and Bell (1997) suggest choosing the threshold where sensitivity equals sensitivity, in other words, where positive and negative observations have equal chance of being correctly predicted. Alternatively, Manel et al. (2001) and Hernandez et al. (2006) maximize the sum of sensitivity and specificity for threshold selection.

Allouche et al. (2006) subtract a constant of 1 from the sum of sensitivity and specificity. This is equivalent to the true positive rate (the proportion of observed presences correctly predicted) minus the false positive rate (the proportion of observed absences incorrectly predicted). They refer to this as the true skill statistic (TSS), and recommend it for model evaluation and comparison, especially when comparing across populations with differing prevalence. In the medical literature, TSS is referred to as Youden's index, and is used in the evaluation of diagnostic tests (Biggerstaff, 2000). However, there is a difference between assessing model performance, and selecting an optimal threshold, and while the true skill statistic itself is independent of prevalence, Manel et al. (2001) found that selecting a threshold to maximize the sum of sensitivity and specificity affects the predicted prevalence of the map, causing the distribution of rare species to be overestimated.

Another way to utilize sensitivity and specificity in threshold selection is to deliberately pick a threshold that will meet a given management requirement. Fielding and Bell (1997) discuss the possibility that a user may have a predetermined required sensitivity or specificity. Perhaps to meet management goals, it is determined that 15% is the minimum acceptable error in the observed presences, and thus a map is required that has a sensitivity of at least 0.85. In a similar vein, Wilson et al. (2005) studied the effects of the various methods of utilizing a probability surface with the goal of defining reserve networks to protect biodiversity. In this work, they contrasted three methods of threshold selection, one of which involved trading off sensitivity to meet a predetermined specificity requirement. They also looked at two methods of working directly from the probability surface, without first creating classification maps.

Another accuracy measure, the kappa statistic, measures the proportion of correctly classified locations after accounting for the probability of chance agreement. While still requiring a choice of threshold, kappa is more resistant to prevalence than PCC, sensitivity and specificity, and was found by Manel et al. (2001) to be well correlated with the area under the curve of ROC plots. Caution is required when using the kappa statistic to compare models across multiple populations. A particular value of kappa from one population is not necessarily comparable to the same kappa value from a different species or location, if the prevalence differs between the two populations (McPherson et al., 2004; Vaughan and Ormerod, 2005; Allouche et al., 2006). Kappa has been used extensively in map accuracy work (Congalton, 1991), and in presence–absence mapping, a threshold can be deliberately selected to maximize kappa (Guisan and Hofer, 2003; Hirzel et al., 2006; Moisen et al., 2006).

While threshold-dependent accuracy measures such as PCC, sensitivity, and specificity have a long history of use in ecology, ROC plots are a technique that has recently been introduced into ecology that provides a threshold-independent method of evaluating the performance of presence–absence models. In a ROC plot the true positive rate (sensitivity) is plotted against the false positive rate (1 − specificity) as the threshold varies from 0 to 1. A good model will achieve a high true positive rate while the false positive rate is still relatively small; thus the ROC plot will rise steeply at the origin, and then level off at a value near the maximum of 1. The ROC plot for a poor model (whose predictive ability is the equivalent of random assignment) will lie near the diagonal, where the true positive rate equals the false positive rate for all thresholds. Thus the area under the ROC curve is a good measure of overall model performance, with good models having an AUC near 1, while poor models have an AUC near 0.5

ROC plots can also be used to select thresholds. As the upper left corner of the ROC plot can be considered the 'ideal' model (sensitivity and specificity both equal 1.0), the threshold which minimizes the distance between the ROC plot and this 'ideal' point can be used as an optimization criteria. In the medical literature, Cantor et al. (1999) performed a review

Table 1 – Number of plots occupied by and prevalence of each of the 13 most common tree species in Zone 16

Latin name	Symbol	Common name	Test plots w/species present	Total plots w/species present
Abies concolor	ABCO	White fir	44	233
Abies lasiocarpa	ABLA	Subalpine fir	72	429
Acer grandidentatum	ACGR3	Bigtooth maple	22	119
Cercocarpus ledifolius	CELE3	Curlleaf mountain-mahogany	29	147
Juniperus osteosperma	JUOS	Utah juniper	103	473
Juniperus scopulorum	JUSC2	Rocky Mountain juniper	45	230
Pinus contorta	PICO	Lodgepole pine	44	230
Pinus edulis	PIED	Common or twoneedle pinyon	92	405
Picea engelmannii	PIEN	Englemann spruce	53	357
Pinus ponderosa	PIPO	Ponderosa pine	38	173
Populus tremuloides	POTR5	Quaking aspen	114	623
Pseudotsuga menziesii	PSME	Douglas-fir	87	417
Quercus gambelii	QUGA	Gambel oak	57	273
Total number of forested plots			386	1930
Total number of plots				2997

of studies that used ROC curve analysis of diagnostic tests. Reviewed studies included several criteria based on the ROC plot, as well as criteria based on all three of the above sensitivity and specificity measures of accuracy. Greiner et al. (2000) review the use of ROC plots in veterinary medicine and discus the use of Youden's index, user required sensitivity and specificity, and ROC plot-based criteria. Biggerstaff (2000), however, demonstrated graphically that in certain circumstances, minimizing this distance can result in a threshold that is inferior both in the probability that a positive prediction is observed present, as well in the probability that a predicted negative is observed absent.

Finally, in many applications, it is important that the mapped prevalence reflects the true prevalence, and thus prevalence itself, both predicted and observed has been used as criteria for threshold selection (Cramer, 2003).

There are numerous criteria by which to choose a threshold to convert a probability surface to a binary map. Here we examine the effect of 11 of these criteria on preserving prevalence, prediction accuracy, and the resulting map output. Comparisons are made using output from presence–absence models developed for 13 tree species in the northern mountains of Utah.

2. Methods

2.1. Data description

Data for these analyses were taken from Moisen et al. (2006) and more detailed information can be found in this paper. In brief, presence of 13 tree species (Table 1) in the northern mountains of Utah were modelled as functions of satellite imagery and topographic information using three different modelling techniques. Of the 3456 plots available in the study area, only forested and single-condition plots were used in these analyses. A total of 1930 sample plots remained, of which 80% (1544 plots) were used for model training. This left 20% (386 plots) for threshold optimization and testing. Analysis was limited to the 13 tree species that were observed present on at least 100 of the 1930 sample plots. Prevalence of these 13 species varied from 0.06 to 0.32, while model quality, as judged by AUC, varied from 0.72 to 0.97 (Table 2).

In this current paper we investigate predictions from one of the modelling techniques (a variant on classification and regression trees implemented in Rulequest's©See5 package) in further detail. We use the probability surface generated by the See5 models for the 13 tree species to compare the utility of different threshold optimization criteria for various map applications.

2.2. Software

Analysis was conducted in the R language and environment for statistical computing (R Development Core Team, 2006). The R package 'PresenceAbsence' (Freeman, 2007) was used

Table 2 – Overall prevalence (from all 1930 plots), AUC (area under the curve from the ROC plot), and range of kappa values for the 13 species

Species	Total prevalence	AUC	Lowest kappa	Highest kappa
ABCO	0.12	0.86	0.26	0.48
ABLA	0.22	0.85	0.31	0.51
ACGR3	0.06	0.87	0.11	0.33
CELE3	0.08	0.82	0.19	0.40
JUOS	0.25	0.96	0.72	0.75
JUSC2	0.12	0.72	0.07	0.19
PICO	0.12	0.97	0.59	0.71
PIED	0.21	0.91	0.56	0.62
PIEN	0.18	0.94	0.46	0.67
PIPO	0.09	0.89	0.30	0.52
POTR5	0.32	0.90	0.57	0.66
PSME	0.22	0.81	0.20	0.43
QUGA	0.14	0.85	0.34	0.48

Area under the curve provides a threshold-independent measure of model quality. The lowest and highest kappa values from the nine criteria (excluding the two user specified criteria) illustrate the sensitivity of each species to choice of criteria. Species with higher model quality (as judged by AUC) and with prevalence closer to 0.5 tend to be less sensitive to choice of criteria.

to optimize thresholds and to produce graphics of the error statistics. This package is available from the CRAN library (http://cran.r-project.org/). Freeman and Moisen (2008) provides a case study and sample code, illustrating the use of this software.

2.3. Threshold criteria

Here we compare 11 threshold optimization criteria for converting a probability surface into a presence–absence map:

(1) *Default*: The traditional default method of setting 'threshold = 0.5'.

(2) *Sens = Spec*: The threshold where sensitivity equals specificity. In other words, find the threshold where positive observations are just as likely to be wrong as negative observations.

(3) *MaxSens + Spec*: The threshold that maximizes the sum of sensitivity and specificity: *Max*(*sensitivity* + *specificity*). In other words, it minimizes the mean of the error rate for positive observations and the error rate for negative observations. This is equivalent to finding the point on the ROC curve whose tangent has a slope of one (Cantor et al., 1999) This is also equivalent to maximizing (sensitivity + specificity − 1), otherwise know as the Youden's index, or the true skill statistic.

(4) *MaxKappa*: The threshold that results in the maximum value of kappa.

(5) *MaxPCC*: The threshold that results in the maximum percent correctly classified.

(6) *PredPrev = Obs*: The threshold where the predicted prevalence is equal to the observed prevalence. In this case we used the observed prevalence from the full data set (1930 plots).

(7) *ObsPrev*: Set the threshold to the observed prevalence. Again, we used the observed prevalence from the full data set (1930 plots).

(8) *MeanProb*: Set the threshold to the mean probability of occurrence from the model predictions.

(9) *MinROCdist*: The threshold that minimizes the straight line distance between the ROC plot and the upper left corner of the unit square, minimizing:

$$(1-\text{sensitivity})^2+(\text{specificity}-1)^2$$

(10) *ReqSens*: The threshold that will give the highest possible specificity, while still meeting a user defined required sensitivity. In other words, the user can decide that the model must misidentify no more than, say, 15% of the plots where the species is observed to be present. Therefore it requires a sensitivity of at least 0.85. This method is useful if, for example, the goal is to define a management area for a rare species, and it is required that the management area does not miss populations.

(11) *ReqSpec*: The threshold that will give the highest possible sensitivity, while still meeting a user defined required specificity. In other words, the user can decide that the model must misidentify no more than, say, 15% of the plots where the species is observed to be absent. Therefore it requires a specificity of at least 0.85. This method is useful if, for example, the goal is to determine if a species is threatened, and it is required that the population is not over-estimated due to true absences misidentified as presences.

The behavior of these last two criteria depends on user specified requirements, and thus these two criteria are not compared directly to the others when assessing criteria. Note that user requirements must be specified with care. If the model is poor, and the requirement is too strict, it is possible that the only way to meet it will be by declaring every single plot to be present (for *ReqSens*) or absent (for *ReqSpec*), resulting in a map with little practical value.

2.4. Criteria assessment

As the error structure can vary between a training data set and independent test data, we used only the test set (386 plots) to optimize the threshold criteria. In this paper we are comparing threshold criteria, and thus require an additional test set to carry out the comparison. Because of small sample size we used cross-validation for this purpose.

For each of the 13 species, fivefold cross-validation was applied to these 386 plots, where four-fifths of the plots were used for threshold optimization by the 11 optimization criteria, and the remaining one-fifth was used for evaluating the resulting classification maps. Kappa and predicted prevalence were used for the evaluation.

The kappa values resulting from the five cross-folds were averaged. Next, star charts were produced for the criteria, with each species represented by one ray on the stars. Criteria with consistently high kappa within most species have stars with a large surface area. Criteria that resulted in the lowest kappa values within multiple species have stars with little surface area.

The rays representing kappa values were rescaled within individual species, so that the criteria with the highest kappa value for a species has a ray for that species of length one, while the criteria with the lowest kappa value has a ray for that species of length zero. By rescaling the range of kappa values independently within each species, we avoided the issues that can be associated with comparing raw kappa values across populations of differing prevalence.

The prevalence bias was calculated by subtracting the observed prevalence for each species (from the 386 test plots) from the average predicted prevalence of the five cross-folds. We examined the prevalence bias by plotting the observed prevalence verses the predicted prevalence from the 13 species for each criteria.

To examine in greater detail the effects of species prevalence and model quality on the threshold criteria, we produced several types of presence absence assessment plots for three species with similar prevalence but varying model quality: JUSC2, ABCO, PICO, and three species with similar model quality, but varying prevalence: PIPO, PIED, POTR5. For final classification mapping (after model validation), Fielding and Bell (1997) recommend maximizing sample size. Thus these final maps and graphs were produced with thresholds optimized on the entire test set (386 plots) rather than on the individual cross-validations.

Finally, to show the spatial effects of threshold criteria, we produced maps comparing the observed presence and predicted presence resulting from selected criteria for the species JUSC2.

3. Results

3.1. *Criteria assessment in terms of kappa*

The threshold criteria that resulted in the highest average kappa for the most species were as follows: *Default*, *MaxKappa*, *PredPrev* = *Obs*, and *MaxPCC*. However, *Default* and *MaxPCC* also resulted in low kappa for many species. In contrast *MaxKappa* and *PredPrev* = *Obs* rarely resulted in the lowest kappa (Fig. 1). *MaxKappa* was amongst the top three criteria for all but one species, and never ranked below 5. *PredPrev* = *Obs* was in the top four criteria for 10 of the 13 species. In addition, when *MaxKappa* and *PredPrev* = *Obs* did not have the highest ranking kappa values, it was often in species such as PIED and JUOS where there was little variation in the optimized thresholds amongst the criteria (Table 2). In contrast, the *Default* threshold was inconsistent, sometimes resulting in high kappa, but in other species performing poorly, leading to low kappa.

In species with low model quality (as judged by AUC) or with low prevalence, the difference between the criteria was the most marked, with some criteria performing well, resulting in high kappa, while other criteria resulted in low kappa. On the other hand, as model quality improved, or prevalence approached 50%, the criteria tend to converge, resulting in similar optimized thresholds and thus similar kappa. For example, species such as PIED (AUC = 0.91, prevalence = 0.21) and JUOS (AUC = 0.96, prevalence = 0.25) have much less variation in their mean kappa than the species with lower model quality or lower prevalence (Table 2).

3.2. *Criteria assessment in terms of preserving prevalence*

The threshold criteria with the lowest bias in the predicted prevalence were *PredPrev* = *Obs* and *MaxKappa*. These were followed by *Default* and *MaxPCC*, both of which slightly under predicted the observed prevalence for most species. The remaining threshold criteria all over predicted the observed prevalence for all 13 species. Note that all 13 species had relatively low prevalence, with the highest being POTR5 with a prevalence of 0.32. If there had been species with prevalence's greater than 0.5, these threshold criteria may have reversed their bias, and under predicted the observed prevalence. Also, the bias of *ReqSens* and *ReqSpec* will vary depending on the user specified requirements (Fig. 2).

3.3. *Effects of criteria choice on particular species*

We begin by examining a histogram of the predicted probabilities, with 'absent' plots represented by light gray and 'present' plots by dark gray (Fig. 3, row 1). For ABCO and PICO the zero bar was truncated (marked by diagonal cross-hatching, with the total given above the shortened bar) to allow the detail on the rest of the plot to be discerned. Notice the double humped histogram for PICO, the species with the highest AUC. An ideal model will result in such a histogram, with all the absent plots on the left, and all the present plots on the right, with no overlap.

Thresholds optimized by three criteria (*Default*, *MaxKappa*, and *PredPrev* = *Obs*) are marked on each plot. As described earlier, species whose models had low AUC (such as JUSC2) have considerable variation between the criteria, however for species with high quality model (such as PICO) the thresholds from each criteria tend to converge.

In Fig. 3 (row 2) the error statistics kappa, sensitivity, and specificity are graphed as a function of threshold. The species with higher model quality (as judged by AUC), maximum kappa increases, and the kappa line flattens resulting in higher kappa values becoming less sensitive to threshold choice. Similarly, for the high quality models, sensitivity and specificity become flatter and higher for a greater range of thresholds, resulting in higher accuracy for both measures at the point where they cross. Finally, as seen in the histograms, as model quality increases, the criteria converge, resulting in greater similarity between the optimal thresholds.

Fig. 3 (row 3) shows the ROC plots. As model quality increases the ROC plot approaches the upper left corner. Not only do the optimized thresholds themselves converge, but the points along the ROC plot that represent these thresholds also converge. This may seem obvious, but the distance along the ROC curve is not uniform. For example, the length of curve between thresholds of 0.1 and 0.2 is not necessarily the same as the length between 0.8 and 0.9.

In Fig. 4, the observed presence–absence data is mapped for species JUSC2. The predicted presence–absence data is also mapped, for thresholds optimized by three criteria: *Default*, *MaxKappa*, and *PredPrev* = *Obs*. JUSC2 is a species with low prevalence (0.12 prevalence) and moderate to low model quality (0.72 AUC). The default criteria (where the threshold is set to 0.5) works particularly poorly for species with low prevalence, drastically underestimating the true occurrence, and producing many false negatives. As one would expect, maximizing kappa provides the highest individual plot accuracy (as judged by kappa), though in JUSC2 it slightly overestimates the true prevalence, producing false positives. Choosing the threshold so that the predicted prevalence equals the observed prevalence will inherently preserve the true occurrence, but results in a slightly lower individual plot accuracy (as judged by kappa). Thus the number of false positives balances the number of false negatives, but the kappa accuracy is slightly lower.

Finally, we look at the special cases of user required accuracy (*ReqSens* and *ReqSpec*), again examining the three species with similar prevalence but increasing model quality. When using the threshold criteria based on user specified accuracy requirements, an optimized threshold is chosen to meet the given requirement (for either sensitivity or specificity), while simultaneously giving the best possible value of the opposing error statistic.

Note that with a very poor model or a very strict requirement, the only way to meet the user requirements may be to set the threshold to 0 or to 1, and declare the species to be present everywhere or nowhere, resulting in maps with little practical value. With good models or low requirements, where

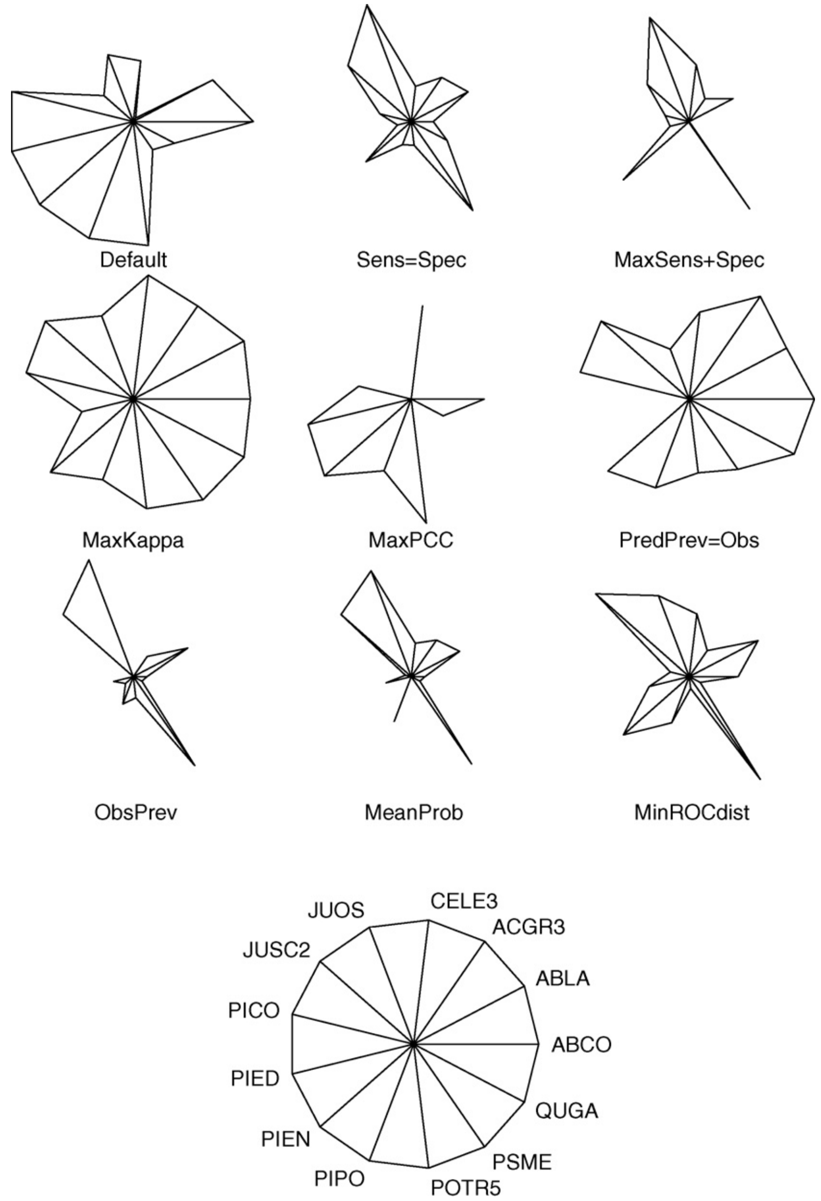


Fig. 1 – Star chart of kappa values of 13 species for each of nine optimization criteria (excluding the two user specified criteria). The star rays are scaled within each species, so that the criteria with the highest kappa for that species has a ray of length 1 and the criteria with the lowest kappa has a ray of length 0. The unit star on the bottom indicates relative positions of species rays in the preceding stars. The criteria *MaxKappa* consistently resulted in the high kappa values. For most species *PredPrev = Obs* also resulted in high kappa values. *Default* performed inconsistently, resulting in the high kappa for some species, but low kappa for other species.

the sensitivity and specificity lines cross at accuracy greater than the user requirements, the unspecified error statistic will have a higher accuracy than the one that was specified in the requirement, and map makers may wish to consider raising their requirements.

We also examined the effects of varying the required specificity within a single species. Fig. 5 shows the histogram, error statistics and ROC plots for JUSC2 with thresholds from four different user required specificities. Fig. 6 shows the resulting predicted presence–absence maps. Specificity is the proportion of observed absences that have been correctly predicted. As the specificity requirement becomes stricter, the threshold rises resulting in fewer plots predicted as presences. Thus, the predicted range of the species also decreases and the number of false positives (observed absences incorrectly predicted as present) decreases. However, the number of false negatives (observed presences incorrectly predicted as absent) rises, and thus the sensitivity decreases.

4. Discussion

Modelling techniques often result in a predicted probability surface, which is then translated into a presence–absence

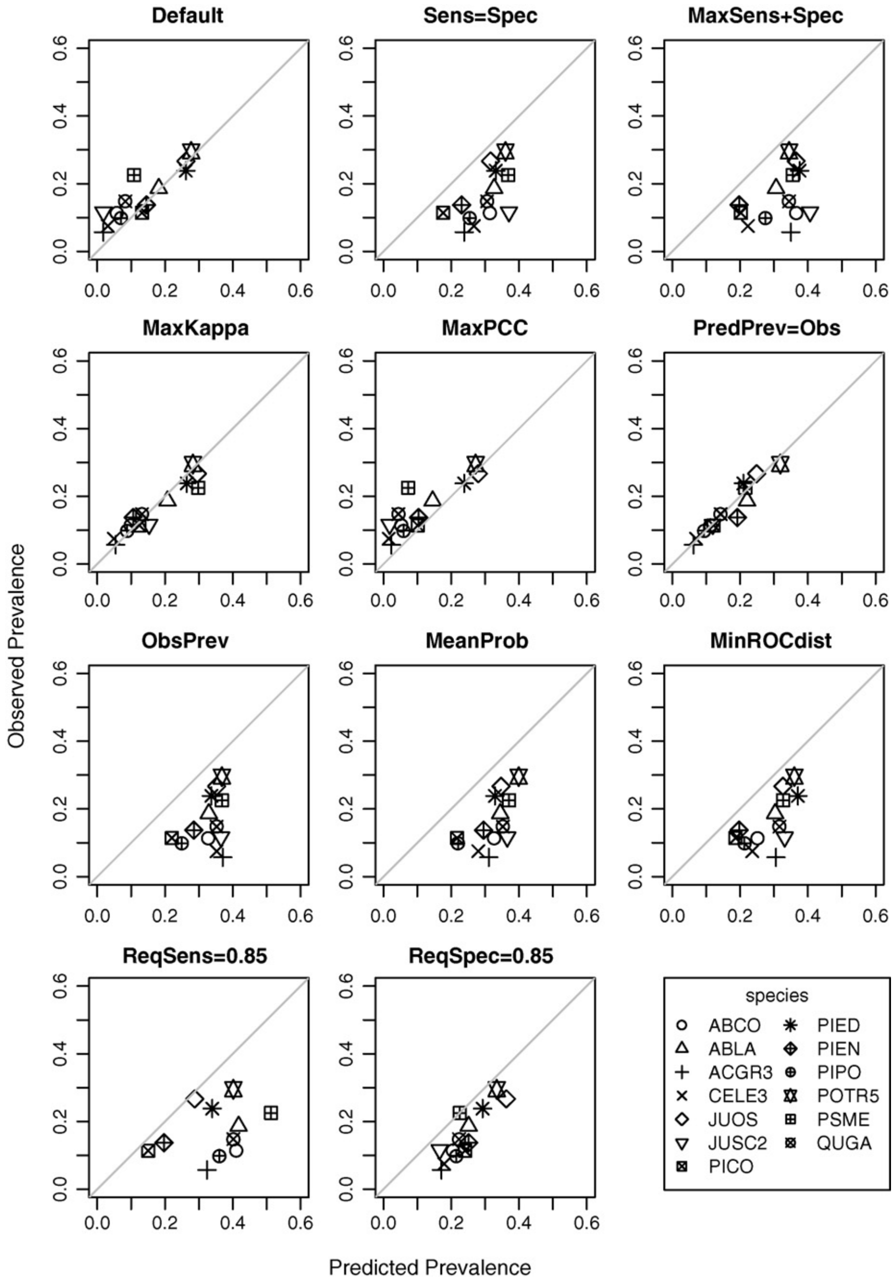


Fig. 2 – Predicted prevalence as a function of observed prevalence. The average predicted prevalence for the five cross-folds is plotted on the *x*-axis and the observed prevalence from the 386 test plots is plotted on the *y*-axis. Threshold criteria with unbiased prevalence have graphs that are symmetric about the diagonal. Criteria that tend to under predict prevalence have graphs with most of the species above the diagonal. Criteria that tend to over predict prevalence have graphs below the diagonal. Note that the graphs for the user specified requirements (*ReqSens* and *ReqSpec*) will vary based on the particular user specifications.

classification map. However, this translation not only requires a (possibly subjective) choice of threshold, it also reduces the information available to the map user. Model quality can be evaluated in terms of both discriminatory ability, and model calibration. Discriminatory ability consists of a models ability to correctly predict presences and absences. Model calibration, on the other hand, concerns the accuracy of the predicted probabilities, in other words, do 40% of the locations assigned a probability of 0.4 actually have the species present. Discrimination and calibration both have ecological implications, and their relative importance depends on the intended use of the model. Pearce and Ferrier (2000) and Vaughan and Ormerod (2005) discuss the implications of model calibration and methods of evaluation model calibration in greater detail. Once the probability surface is translated to a classification map, it is only possible to evaluate it in terms of discrimination, limiting the possible insight into the model.

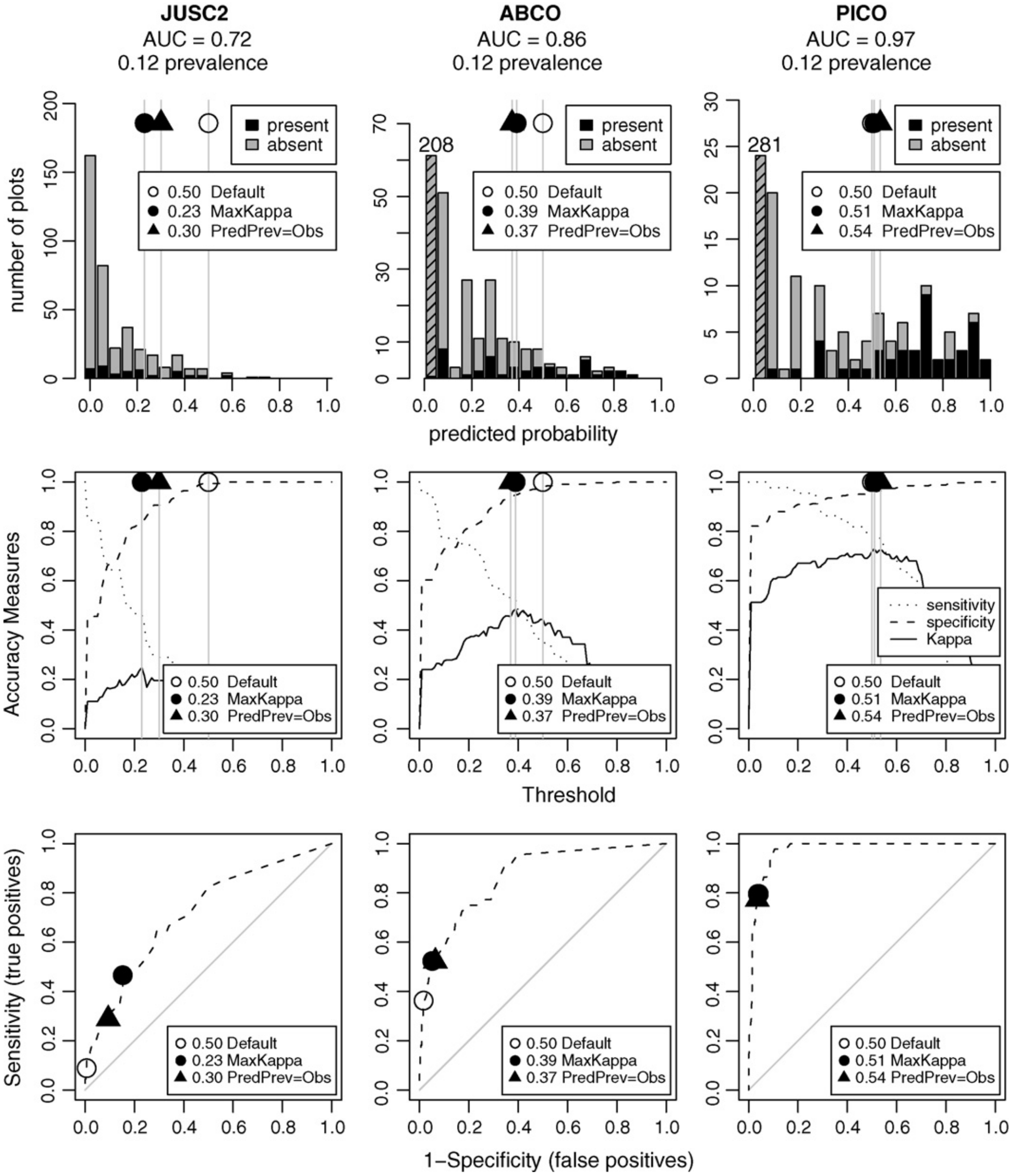


Fig. 3 – Histogram, accuracy measures, and ROC plots for three species with similar prevalence, but increasing model quality. Histogram bars are broken into plots that were observed present (dark gray) and plots that were observed absent (light gray). For ABCO and PICO, the zero bars are truncated to fit on plot (indicated by cross-hatching) and the actual height given above the bar. Optimized thresholds are marked for the two criteria with the highest kappa and the most accurate predicted prevalence from the cross-validation. The standard default threshold of 0.5 is also indicated.

In addition, one of the most powerful techniques for evaluating discriminatory ability is ROC plots with their associated AUC, which requires the full probability surface to calculate. ROC plots are independent of both threshold and prevalence, allowing the comparison of model discriminatory ability across species and locations.

Results from these analyses have implications for species mapping efforts. Threshold cut-offs should be chosen in light of the intended use of the species distribution maps. In most mapping applications for forest management, 0.5 is chosen as the default threshold cut-off. Yet, analyses here illustrate that for species with low prevalence or low model quality, a 0.5 cut-off is unreliable, sometimes resulting in substantially lower kappa, with possible detrimental effects on a management decision.

If one's goal is to accurately predict the overall prevalence of rare species, be aware that many of the threshold criteria proposed in the literature also resulted in maps that substantially overestimate the range of low prevalence species. While the AUC and ROC plots are prevalence-independent, threshold choices based on them are not, nor are choices based on other criteria that optimize statistics based on sensitivity and specificity. These include *MinROCdist*, *Sens = Spec*, and *MaxSens + Spec*. Thresholds set in this manner tend to overestimate the prevalence of rare species while underestimating the prevalence of common species. This fact was noted by Manel

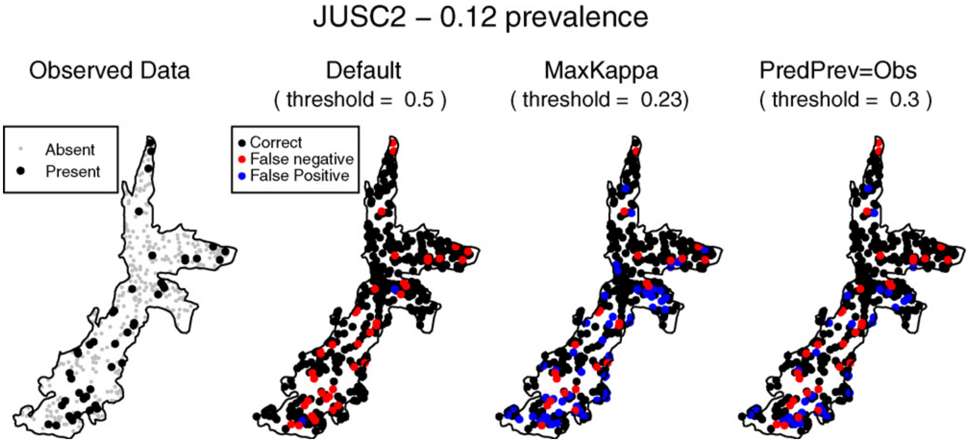


Fig. 4 – Observed presence absence data for species JUSC2, from 386 test plots in zone 16, and presence absence predictions for the same plots with thresholds chosen by three different criteria: the traditional default criteria of 0.5 (*Default*), the threshold that maximizes kappa accuracy (*MaxKappa*), and the threshold that best preserves the prevalence of the species (*PredPrev = Obs*).

et al. (2001), and our analyses support it. Therefore, for map making, if preserving prevalence is important to the map's intended use, we found selecting a threshold to maximize kappa to be preferable to maximizing the sum of sensitivity and specificity.

The *Default* threshold of 0.5 is known to under predict the prevalence of rare species, and we found this to be true for our data, particularly when low prevalence was accompanied by low model quality. It has been suggested that choosing a threshold to equal prevalence (*ObsPrev*), or to equal the mean predicted probability (*MeanProb*) may be a preferable alternative to using the *Default* threshold of 0.5 when dealing with species with low prevalence (Cramer, 2003). However we found that replacing the *Default* criteria of 0.5 with *ObsPrev* or *MeanProb* did not improve kappa, and, in effect, merely exchanged over predictions for under predictions.

If a management objective above all requires a map to portray unbiased estimates of species prevalence, then the best results were obtained from thresholds deliberately chosen so that the predicted prevalence equaled the observed prevalence (*PredPrev = Obs*), followed closely by thresholds chosen to maximize kappa (*MaxKappa*). Fortunately, these were also the two criteria with the highest mean kappa on the independent test data.

It may seem self-evident that a threshold chosen to maximize kappa will result in the highest kappa, and a threshold chosen to preserve prevalence will result in the most accurate prevalence; however, more importantly, we also found the criteria that maximized kappa was the second best at preserving prevalence, and the criteria that preserved prevalence was the second best at maximizing kappa.

We also found that species with poor model quality, or low prevalence were most sensitive to the choice of threshold. If a species has a good quality model, and prevalence near 50%, then any optimization criteria, including the traditional default method of 0.5, may result in equally useful maps.

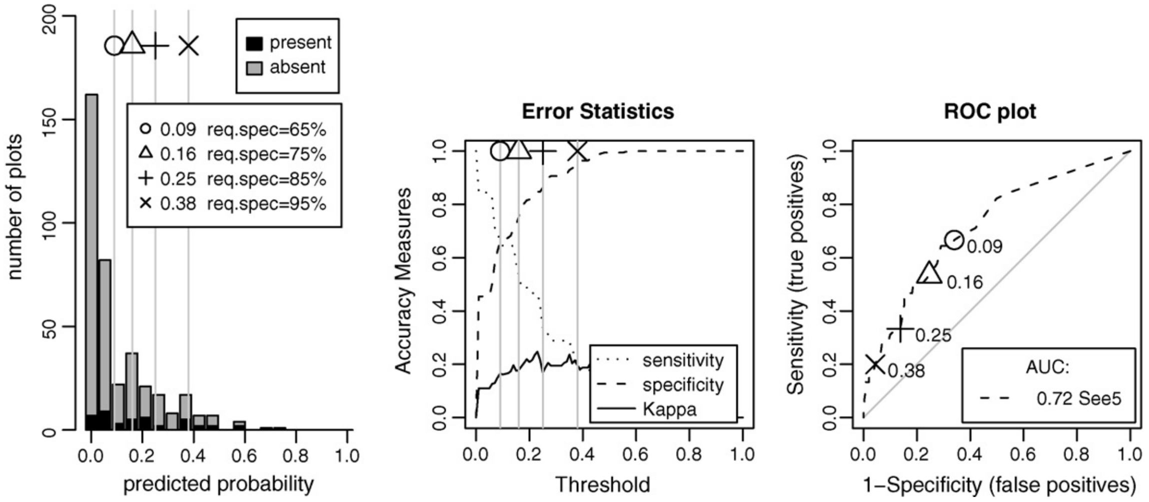


Fig. 5 – Accuracy measures for species JUSC2, using the user specified *ReqSpec* criteria, for four different required levels of specificity.

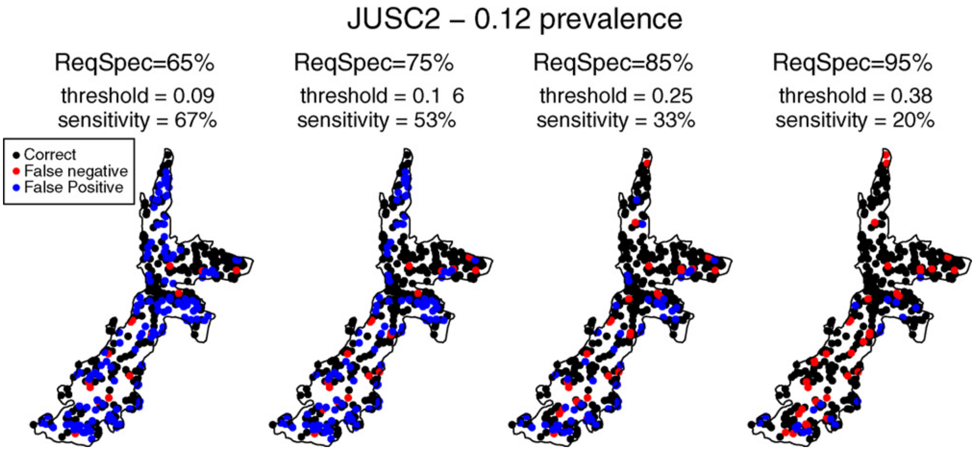


Fig. 6 – Predicted presence absence data for JUSC2 using the user specified *ReqSpec* criteria, for four different required levels of specificity. As the specificity requirement rises, the resulting sensitivity decreases. In other words, there is a trade-off between false positive predictions and false negative predictions.

However, for species with poor model quality or low prevalence, matching the threshold criteria to the maps intended use becomes more important.

For particular management applications the special cases of user specified required accuracy (*ReqSens* and *ReqSpec*) may be most appropriate. For example, if the goal is to determine if a species is threatened, and it is important to avoid over-inflating the population by misclassifying true absences as predicted presences, then a user-defined required specificity (*ReqSpec*) may be best. In this case, the user may decide that it is unacceptable to classify more than, for example, 5% of the true absences as present, and thus they require a map with a specificity of 0.95.

Conversely, if a map is to be used as a pre-stratification to narrow the search window for a particular species in a sampling effort, and it is imperative that the field survey include all potential habitats, then a user-defined required sensitivity (*ReqSens*) may be the best approach. In this case, the user may decide that it unacceptable to miss more than, for example, 1% of the true locations, and thus they require a map with a sensitivity of 0.99.

If a map has a high specificity (and thus a low sensitivity) one can be very confident that anywhere mapped as present actually does have the species. However, one has much less confidence in the areas that have been mapped as absent. They may be true absences, but there is a substantial chance that they may be true presences that have been misclassified. The converse is true for a map with high sensitivity.

Ultimately, maps will typically have multiple and somewhat conflicting management applications and thus providing users with a continuous probability surface may be the most versatile method, not only allowing threshold choice to be matched with map use, but also allowing the users to distinguish between a map's discrimination and its calibration. Model evaluation can be carried out on this probability surface, rather than on particular classification maps. The AUC, which requires the full probability surface to calculate, provides an objective measure of a model's discrimination ability, not dependent on threshold choice or the prevalence of the population. In addition, providing users with the probability surface allows the examination of the models calibration, which can be critical to some ecological applications, and is impossible to determine from a classification map. Finally, threshold choice can be matched to each map's use, allowing the maps to become a more powerful management tool.

The authors thus suggest that mapmakers produce continuous probability maps rather than binary species distribution maps enabling the map user to choose appropriate threshold cut-off values in light of the intended map use. This is particularly critical for low prevalence species, or for models with lower discriminatory ability, where the choice of optimization criteria has a more dramatic effect on threshold. Hopefully this paper has presented a set of tools to help map users work with probability surfaces, and create classification maps to meet their particular management needs.

REFERENCES

Allouche, O., Tsoar, A., Kadmon, R., 2006. Assessing the accuracy of species distribution models: prevalence, kappa, and the true skill statistic (TSS). Journal of Applied Ecology 43 (6), 1223–1232.

Biggerstaff, B.J., 2000. Comparing diagnostic tests: a simple graphic using likelihood ratios. Statistics in Medicine 19 (5), 649–663.

Cantor, S.B., Sun, C.C., Tortolero-Luna, G., Richards-Kortum, R., Follen, M., 1999. A comparison of C/B ratios from studies using receiver operating characteristic curve analysis. Journal of Clinical Epidemiology 52 (9), 885–892.

Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment 37 (1), 35–46.

Cramer, J.S., 2003. Logit Models from Economics and Other Fields. Cambridge University Press, pp. 66–72.

Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24 (1), 38–49.

Freeman, E., 2007. PresenceAbsence: An R Package for Presence–Absence Model Evaluation. USDA Forest Service, Rocky Mountain Research Station, 507 25th street, Ogden, UT, USA. http://cran.r-project.org/.

Freeman, E., Moisen, G., 2008. PresenceAbsence: an R package for presence–absence analysis. Journal of Statistical Software 23, 11.

Greiner, M., Pfeiffer, D., Smith, R.D., 2000. Principles and practical application of the receiver-operation characteristic analysis for diagnostic tests. Preventive Veterinary Medicine 45, 23–41.

Guisan, A., Hofer, U., 2003. Predicting reptile distributions at the mesoscale: relation to climate and topography. Journal of Biogeography 30, 1233–1243.

Hernandez, P.A., Graham, C.H., Master, L.L., Albert, D.L., 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29, 773–785.

Hirzel, A.H., Le Lay, G., Helfer, V., Randin, C., Guisan, A., 2006. Evaluating the ability of habitat suitability models to predict species presences. Ecological Modelling 199, 142–152.

Manel, S., Williams, H.C., Ormerod, S.J., 2001. Evaluating presence–absence models in ecology: the need to account for prevalence. Journal of Applied Ecology 38, 921–931.

McPherson, J.M., Jetz, W., Rogers, D.J., 2004. The effects of species range sizes on the accuracy of distribution models: ecological phenomenon or statistical artifact. Journal of Applied Ecology 41, 811–823.

Moisen, G.G., Freeman, E.A., Blackard, J.A., Frescino, T.S., Zimmerman, N.E., Edwards, T.C., 2006. Predicting tree species presence and basal area in Utah: a comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. Ecological Modelling 199, 176–187.

Pearce, J., Ferrier, S., 2000. Evaluating the predicting performance of habitat models developed using logistic regression. Ecological Modelling 133, 225–245.

R Development Core Team, 2006. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN: 3-900051-07-0. http://www.R-project.org/.

Vaughan, I.P., Ormerod, S.J., 2005. The continuing challenges of testing species distribution models. Journal of Applied Ecology 42, 720–730.

Wilson, K.A., Westphal, M.I., Possingham, H.P., Elith, J., 2005. Sensitivity of conservation planning to different approaches to using predicted species distribution data. Biological Conservation 22 (1), 99–112.