Research

Ecography 35: 250–258, 2012
doi: 10.1111/j.1600-0587.2011.06545.x

Subject Editor: Carsten Rahbek. Accepted 24 May 2011

The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models

William T. Bean, Robert Stafford and Justin S. Brashares

W. T. Bean (bean@berkeley.edu) and J. S. Brashares, Dept of Environmental Science, Policy and Management 6, Univ. of California, Berkeley, CA 94720, USA. – R. Stafford, California Dept of Fish and Game, PO Box 6360, Los Osos, CA 93402, USA.

Species distribution models are used for a range of ecological and evolutionary questions, but often are constructed from few and/or biased species occurrence records. Recent work has shown that the presence-only model Maxent performs well with small sample sizes. While the apparent accuracy of such models with small samples has been studied, less emphasis has been placed on the effect of small or biased species records on the secondary modeling steps, specifically accuracy assessment and threshold selection, particularly with profile (presence-only) modeling techniques. When testing the effects of small sample sizes on distribution models, accuracy assessment has generally been conducted with complete species occurrence data, rather than similarly limited (e.g. few or biased) test data. Likewise, selection of a probability threshold – a selection of probability that classifies a model into discrete areas of presences and absences – has also generally been conducted with complete data. In this study we subsampled distribution data for an endangered rodent across multiple years to assess the effects of different sample sizes and types of bias on threshold selection, and examine the differences between apparent and actual accuracy of the models. Although some previously recommended threshold selection techniques showed little difference in threshold selection, the most commonly used methods performed poorly. Apparent model accuracy calculated from limited data was much higher than true model accuracy, but the true model accuracy was lower than it could have been with a more optimal threshold. That is, models with thresholds and accuracy calculated from biased and limited data had inflated reported accuracy, but were less accurate than they could have been if better data on species distribution were available and an optimal threshold were used.

Species distribution models (aka habitat suitability models or environmental niche models) are used to examine a multitude of evolutionary, ecological and conservation-related questions. Applications of species distribution models (SDMs) range from identifying conservation gaps (Kremen et al. 2008), predicting the effects of climate change (Thuiller 2004), and exploring mechanisms of speciation (Graham et al. 2004). Such models use geolocated species presence records and spatially explicit environmental and ecological correlates to create a map of presence probability. Approaches to species distribution modeling include more traditional regression techniques (e.g. generalized linear models and generalized additive models) to more recent techniques codified in software packages (e.g. Maxent). A thorough review of approaches is provided by Elith et al. (2006), Graham and Hijmans (2006) and Elith and Leathwick (2009).

Given that occurrence data are sparse for most species, small datasets (generally considered <100 occurrence points (Hernandez et al. 2006), due either to species rarity or incomplete historical collections, are increasingly being used in suitability modeling (Gibson et al. 2007, Papes and Gaubert 2007, DeMatteo and Loiselle 2008). The use of imperfect species records (i.e. species records with a small number of data points and/or spatially biased records) has become increasingly popular as modeling techniques have grown more robust. In particular, the species distribution modeling software Maxent (Phillips et al. 2006) is often cited as a particularly good model for small datasets. However, while some work has been done to explore the accuracy of models with small sample sizes (Hernandez et al. 2006), less attention has been paid to the effects imperfect records will have on the two critical secondary steps in modeling methods: threshold selection and accuracy assessment.

Threshold selection

The primary output of most SDMs is a raster representing the probability of species occurrence. For most applications, it is often necessary to select a threshold of probability to classify each pixel into two categories, 'suitable', or 'present', and 'unsuitable' or 'absent'. Each pixel is deemed suitable if the probability of occupancy is greater than the threshold value, and vice versa. Most common techniques for assessing the accuracy of SDMs require this binary map, as do most model applications. For example, identifying areas to target for invasive species mitigation (Ward 2007), mapping areas of connectivity (Waltari and Guralnick 2009), and modeling range expansion (De Marco et al. 2008) all require delineating boundaries of species presence and absence from a map of probability.

The simplest probability threshold classifies all areas of probability greater than 0.5 as 'present' and all areas below 0.5 as absent (Manel et al. 1999). Although this technique intuitively makes the most sense, it has generally been deprecated for a number of statistical reasons including that the ratio of presences to absences in the training model are rarely equal, thus the mean probabilities of each event are biased (Guisan and Theurillat 2000). Another approach sets the probability threshold to the prevalence of modeled data: that is, the ratio of presence points to total observations in the model data (Cramer 2003). More advanced techniques for selecting a probability threshold require analyzing the model's sensitivity and specificity using species occurrence data that is withheld from the model-building process. Sensitivity is defined as the percent of 'true' presences correctly classified as present by the model, and specificity is the percent of 'true' absences correctly classified as absent (Liu et al. 2005).

Other advanced methods use the receiver operating characteristic (ROC) curve to calculate a threshold. The receiver operating characteristic (ROC) curve plots sensitivity against 1 minus specificity. In a completely discriminating model, sensitivity would remain at 1.0 across all levels of specificity, and any threshold selected would be correct, where all presences and absences are correctly classified. Methods for calculating a probability threshold using the ROC include selecting a threshold where specificity equals sensitivity, maximizing the sum of specificity and sensitivity, or selecting the probability corresponding to the point on the ROC curve closest to (0,1) (Liu et al. 2005). In these approaches, withheld records of species occurrence are required in order to test the specificity and sensitivity of the model.

A third category of threshold selection identifies a threshold value that maximizes the percent of points correctly classified (PCC); maximizes sensitivity plus specificity; or maximizes Kappa, a measure that utilizes both sensitivity and specificity (Guisan et al. 1998). Finally, many studies use the lowest presence threshold (LPT) (Pearson et al. 2007, Raxworthy et al. 2007, Brown et al. 2008, Thorn et al. 2009), an approach in which the threshold is set to the lowest probability found at any of the confirmed species presence points. The LPT approach has commonly been modified to include only a certain percentage of presence points. For example, some studies select a threshold that would classify 95% of all known presence points as present (Waltari and Guralnick 2009). Importantly, these techniques require that spatially defined field data on species occurrence data are available to calculate the threshold and test model accuracy. While Maxent may be used with presence-only data, of course only those models with small datasets on presence and absence may be used for these secondary modeling steps.

All of the techniques of selecting a threshold described above have been analyzed across a range of sample sizes and with multiple accuracy metrics to test their ability to produce results that match 'true' occurrences. Studies thus far have relied on relatively large sample sizes, either to create the models (Liu et al. 2005), or to test their accuracy and select thresholds (Hernandez et al. 2006, Jimenez-Valverde and Lobo 2007, Freeman and Moisen 2008a). However, species distribution modeling, threshold selection and accuracy assessment are all components of a single approach. In real world situations, small samples and biased data used to create the model must also be used to calculate a threshold and to assess the model accuracy. Errors in the original data can be compounded throughout the model validation process to create misleading results. Given this, it is surprising that little has been done to examine the synergistic effects of small sample size and/or biased data on model output.

Spatial and temporal bias in field collection

Although the effect of sample size on species distribution modeling has been well explored, sample bias represents another common problem in species presence data, especially for rare species or historic datasets (Phillips et al. 2009). Species' distributions can be biased in space (e.g. biased sampling effort) or time. In the case of time, a species may have been sampled at a time of contraction, representing a much more limited range than seen at times of expansion. While species distribution models have been suggested to be accurate with small samples (Hernandez et al. 2006), and efforts are underway to address spatial bias (Phillips et al. 2009), less is know about how temporal bias may affect modeling, threshold selection and accuracy assessment in combination.

In this study, we used multiple years of population-level distribution data for the giant kangaroo rat (GKR, *Dipodomys ingens*), a federally and state-listed endangered rodent endemic to California, to examine the effects of varying sample size and sample bias on threshold selection and accuracy assessment. By combining several years of range-mapping data we were able to create a baseline distribution that incorporated expansion and contraction over time. We then compared model predictions from this multi-year data set with those based on individual years of collection, and across multiple sample sizes. We did this with two objectives: 1) to compare thresholds calculated from each model of limited data with the 'optimal' threshold calculated from the combined (i.e. multi-year) distribution for that model, and 2) to compare accuracy reported from the limited data with the accuracy of the model calculated from the combined distribution. In doing so, we identified 1) how accurate the models appeared to be, given the threshold and accuracy calculated from the limited data, 2) how accurate the models really were given the threshold calculated from limited data and accuracy calculated from the multi-year (i.e. 'true') distribution, and 3) how accurate they could have been, given a threshold and accuracy calculated from the multi-year distribution (Fig. 1).

Methods

We utilized three years of GKR population distribution data, each representing different levels of sample bias, with subsample sizes of 5, 10, 100 and 1000 to assess the effects of sample size and sample bias on threshold selection and accuracy assessment. For each year and sample size, we tested seven methods of threshold selection and four methods of accuracy assessment.

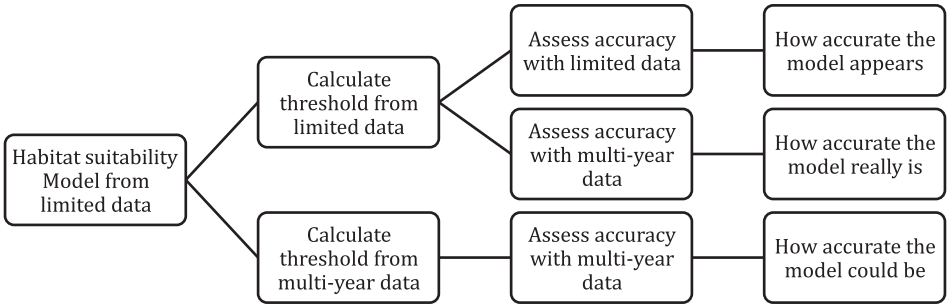

Figure 1. Flow chart for analyzing model output with limited data. Distribution models were created for 2001, 2006, 2008, and the multi-year distribution with sample sizes of 5, 10, 100 and 1000. Each model was evaluated in three ways: how accurate it appeared to be (with threshold and accuracy calculated from the limited data), how accurate it was (with threshold calculated from the limited data and accuracy calculated from the multi-year data) and how accurate the model could be (with threshold and accuracy calculated from the multi-year data).

Study site and focal species

The giant kangaroo rat (GKR) is a burrowing rodent that inhabits dry grassland habitat, clipping the vegetation around their burrows for forage and visibility (Braun 1985). This behavior makes active burrows contrast strikingly with areas without GKRs, with burrows appearing as circles of bare soil 4–6 m in diameter in a matrix of standing grass. In years with a large standing crop of grass, these bare areas, and thus the spatial extent of the population, is easily mapped during aerial surveys.

Once distributed throughout California's San Joaquin Valley, GKR are now limited to a half dozen, fragmented populations in and around the Coast Range (Williams 1992). The largest of these populations occurs in Carrizo Plain National Monument ('Carrizo', Fig. 2). The Carrizo Plain covers approximately 820 km^2 and is considered the largest remnant of the San Joaquin Valley Grassland habitat. Precipitation in Carrizo is low (mean = 20 cm annually) and variable (SD = 10 cm). Such variability is believed to contribute to dramatic annual changes in GKR abundance and distribution. Based on aerial surveys, in the span of only five years, GKRs in Carrizo nearly doubled the size of their distribution. The GKR in Carrizo thus serves as a model species for examining the impact of a variable distribution on SDM accuracy: its large shifts in distribution allow for investigation into multiple, real-world examples of possible sampling bias in both time and space, while multiple years of mapping its areal distribution present a nearly complete measure of their maximum potential distribution. Finally, our ability to map GKR areal distribution through time allows for a comparison between models created with potential small and biased samples with models created from near-complete knowledge of GKR distribution.

Population distribution mapping and model creation

In 2001 and 2006, we conducted aerial transects every quarter mile across all of the Carrizo to map GKR activity throughout the monument (Fig. 2). The distribution in 2001 represented a lower year in GKR distribution, whereas 2006 more closely approximated the full possible extent. In 2006, GKR expanded across most of the plain. Further expansion would have required incursion into steep slopes of oak woodland or other areas dominated by *Avena* grass species, both generally considered uninhabitable by GKR. Thus, it was shown empirically that the distribution of GKR in 2001 was much smaller than the population's full range. In 2008, aerial surveys were only conducted across the northwest portion of the GKR range, representing a clear spatial sample bias. In addition to these data, we created a combined population distribution with the Merge function in ArcGIS 9.2 (ESRI 2008). This combined distribution represented, as near as possible, an empirically-derived estimate of the full potential

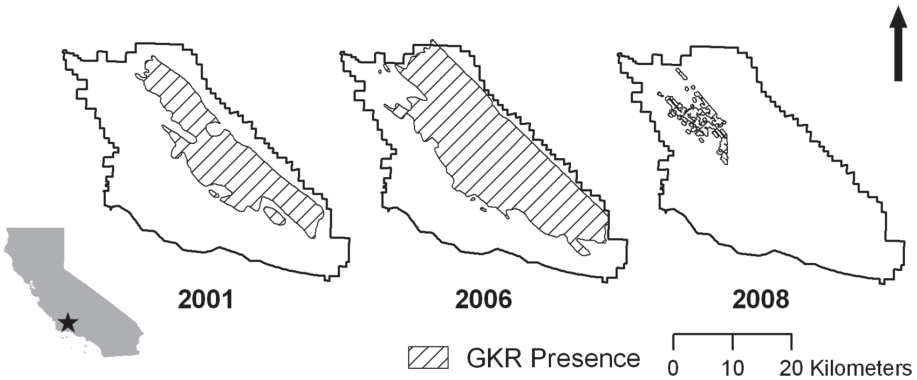

Figure 2. Results of aerial surveys, 2001, 2006 and 2008. Species presence points were randomly sampled from within each of these distributions for the given year. Pseudo-absences were sampled as background points in Maxent. Absence points were randomly sampled from outside of the distribution for each year for model validation. Representative Maxent model results given in Supplementary material Appendix 1.

distribution of giant kangaroo rats in Carrizo, and served as a test of the accuracy of models built using the limited and biased datasets.

Species distribution models of GKR in Carrizo were created for each year separately with the software Maxent (Phillips et al. 2006). Maxent uses a maximum entropy approach to estimate the most uniform distribution of a species' occurrence across the study area, constrained by the provided environmental correlates. Maxent was chosen because it has been shown to work better than other approaches with small samples (Hernandez et al. 2006). Importantly, Maxent incorporates observed presence points in model building, but does not incorporate points where the focal species is known to be absent. Instead, it relies on thousands of background 'pseudo-absence' points picked at random throughout the area of interest to characterize unoccupied habitat. A more complete discussion on how pseudo-absences are selected, and their utility in species distribution modeling can be found in Wisz and Guisan (2009). Because of the use of pseudo-absences, modeled data 'prevalence' (the ratio of presence to total points) tends to be quite low. The default number of background points is 10 000, making prevalence with 50 observations 0.005. In this case, setting the probability threshold to the original prevalence is inadvisable (Jimenez-Valverde and Lobo 2007). In presence-pseudo-absence models, the ratio of presence points to total modeling points is not directly comparable with the probability of occurrence. Pseudo-absences are almost certain to contain species presence points, and thus 'prevalence' calculated from a profile model will be negatively biased.

We created 10 random sets of presence points from within the observed distribution of GKR for each year of data (2001, 2006, 2008, and combined), with sample sizes of 5, 10, 100 and 1000 using the random point generator in the Hawth's tools extension for ArcGIS (Beyer 2004). For each year and sample size, we created 10 models in Maxent, for a total of 160 models. We used soil type at the group level from the SSURGO database (USDA 2008), vegetation type (USDA Forest Service 2008), elevation (USGS 2008), slope (calculated from elevation), and precipitation isoclines (USGS 1994) as explanatory variables, and used the Maxent default feature selection algorithm. Soil, vegetation and precipitation isoclines were obtained as vector polygons, and converted to raster to match the scale of the digital elevation model (DEM). All explanatory variables were calculated at 30 m resolution. The study area consisted of 1 111 365 pixels (~1000 km^2). Giant kangaroo rat extent in 2001 was 295 152 pixels (~265 km^2), 539 280 pixels (485 km^2) in 2006, and 39 310 pixels (35 km^2) in 2008. Mean temperature or variation in temperature showed little variation at the scale of this study and, thus, these variables were not included in the models. All other Maxent parameters were set to the default.

To assess the accuracy of the models, we calculated the area under the curve (AUC), a commonly reported threshold-independent measure of model accuracy (Hanley and McNeil 1982). AUC represents the area under the ROC curve. A perfect model would have an AUC of 1, and a model no different from random would have a score of 0.5. First, we calculated AUC using data comparable in sample size to the model-training data. For example, AUC for a model created from a sample of 5 points from the 2001 distribution was calculated using 5 absence and 5 presence points randomly selected from the 2001 distribution. We also calculated the 'true' AUC using 2000 points randomly selected from the combined distribution as a comparison. In short, this approach considered 1) how accurate the models appeared to be, with accuracy calculated from limited data on GKR presence and absence, and 2) how accurate they were, with accuracy calculated with more complete knowledge of their potential distribution.

Threshold selection

For each of the 160 models, we used seven methods of threshold selection; six recommended by Liu et al. (2005), as well as the most commonly used method, LPT (Table 1). In addition, we selected thresholds based on four levels of required sensitivity: 50, 90, 95 and 99%. In this approach, a separate threshold was calculated to produce each level of desired sensitivity. Such an approach can be useful in conservation planning, for example, when a high, specific level of sensitivity is necessary.

For each model, two thresholds were calculated: first, a threshold based on data with a comparable sample size and from the same distribution as the model-training data. For example, for a model with a sample size of 5, a threshold was calculated from an additional 10 points (5 presence and 5 absence points) randomly selected from the single-year distribution. A second threshold was calculated using 2000 points (1000 presence and 1000 absence) randomly selected from the multi-year distribution. This second threshold represents the best possible threshold for each model assuming the user has knowledge of the potential distribution of the species. That is, given a model constructed with a small sample size and from a single year, but knowing the complete distribution of the species, what is the optimal threshold to classify the model into suitable and non-suitable habitat? The second threshold serves as a comparison to the threshold calculated from the limited data. All thresholds were calculated in R (R Development Core Team 2009) using the PresenceAbsence package (Freeman and Moisen 2008b).

Table 1. The eight distinct approaches used to select threshold values in species distribution models considered in this study. These approaches have been recommended for a variety of taxa, data types, and model approaches.

Method	Approach	Reference
Method 1	Sensitivity = specificity	Cantor et al. 1999
Method 2	Maximize (sensitivity + specificity)	Cantor et al. 1999, Manel et al. 2001
Method 3	Maximize kappa	Huntley et al. 1995
Method 4	Maximize PCC	Guisan et al. 1998
Method 5	Mean probability of model building points	Cramer 2003
Method 6	Closest point to (0,1) on ROC curve	Cantor et al. 1999
Method 7	Lowest presence point probability	Pearson et al. 2007
Method 8	Set to achieve desired sensitivity	

Accuracy assessment

For each model and threshold, we calculated four measures of accuracy: percent of points correctly classified, sensitivity, specificity and Kappa (Cohen 1960). Each accuracy measure was calculated in three ways. First, as with the threshold calculations, we measured accuracy using a sample matched in size to the model-building data (e.g. 100 presence and 100 absence points for an original sample size of 100). We also calculated accuracy using the multi-year distribution. The first measure of accuracy demonstrated how accurate the model appeared to be, given limited knowledge of a species' distribution. The second showed how accurate the model actually was, given near-complete knowledge of its potential distribution in Carrizo. Finally, we calculated the accuracy of the model using the optimal threshold calculated from the multi-year distribution. This measure of accuracy represented how accurate the model could be given an optimal threshold calculated from near-complete knowledge of a species' distribution. Accuracy was measured using the PresenceAbsence package (Freeman and Moisen 2008b).

Results

Population distribution mapping and model creation

The years 2001 and 2006 represented quite different distributions for GKR (Fig. 2). In 2001 there was a smaller population, limited in distribution, whereas in 2006 the population covered nearly all of the Carrizo Plain. In 2008, due to sampling bias, there was a limited distribution and, consequently, a limited distribution model. Maxent results for each sample size and year are provided in Supplementary material Appendix 1.

Threshold selection

Thresholds calculated from limited data tended to be higher than the optimal threshold, particularly in spatially biased samples and with larger sample sizes (Fig. 3 and Supplementary material Appendix 1, Fig. A2). There were few differences among the six methods recommended by past studies (Liu et al. 2005, Jimenez-Valverde and Lobo 2007). The LPT approach created much lower thresholds across all samples and model years. Thresholds calculated for a required level of sensitivity were uniformly higher when calculated from limited data than from the multi-year distribution. For models created from 2008 distribution data, thresholds were presented for sensitivities that were, in reality, impossible to achieve. Due to the limited distribution model, setting the threshold only to 0 (e.g. all pixels classified as present) would have resulted in the desired sensitivity (e.g. 99% points correctly classified as present). In 2001 and 2006, and in the multi-year distribution, thresholds calculated from the limited data were closer to the optimal threshold for the model, especially at larger sample sizes (Fig. 3).

Accuracy assessment

Across all four measures of accuracy and all seven methods of threshold selection, accuracy calculated using the limited data was higher than the accuracy calculated from the multi-year distribution (Fig. 4, Supplementary material Appendix 1, Fig. A3–A7). In other words, limiting the data used to test accuracy created the false conclusion that those models were accurate. However, in most cases, model accuracy was lower than it could have been if the 'optimal' threshold calculated from the multi-year distribution had been used. This pattern was most prominent in 2008, the spatially biased sample, but consistent in 2001 and 2006. The LPT method of threshold selection created models essentially no better than random, with most models classifying only 50% of points correctly. Interestingly, on average, models from 2008 with small sample size appeared to be more accurate than models with a large sample size.

AUC scores for all models calculated from limited data were universally good (Fig. 4). However, AUC calculated from the multi-year distribution revealed some models to be poor performers. At low sample sizes, there were no significant differences in AUC across the different models. On average, the multi-year model did as well as any of the single-year models, including 2008. However, AUC values calculated from limited data made the accuracy of the models appear much higher than it would have been if it had been calculated from more complete knowledge of the potential distribution. In 2008, for example, AUC scores were consistently 15% higher when calculated from limited data than the multi-year distribution. As sample size increased in 2001 and 2006, AUC appeared to stay constant when calculated from the limited data, however, the 'true' AUC increased. Not surprisingly, models with larger sample sizes were more accurate than those with small samples. The differences between AUC calculated from limited and multi-year data decreased as sample size increased.

Reported and true accuracy for the models from the 2006 data and the multi-year distribution converged at 100 points. That is, with reasonably unbiased data, thresholds can be optimally selected and accuracy appropriately assessed with approximately 100 separate model and evaluation points (Fig. 4).

Discussion

Species distribution modeling is a large and growing field in ecology, but data used to develop models frequently suffer from small sample size and/or spatially or temporally biased data collection. Some studies have shown that models created from small samples can be accurate, but many tests of small samples assume knowledge of a species' full distribution to calculate accuracy estimates (Hernandez et al. 2006). This study demonstrates that probability thresholds and accuracy assessments calculated from limited knowledge of a species' distribution compound the problems of small sample size and bias and, perhaps more importantly, may create unfounded confidence in model results.

Problems of small sample size and sample bias were also apparent in selecting an appropriate threshold for

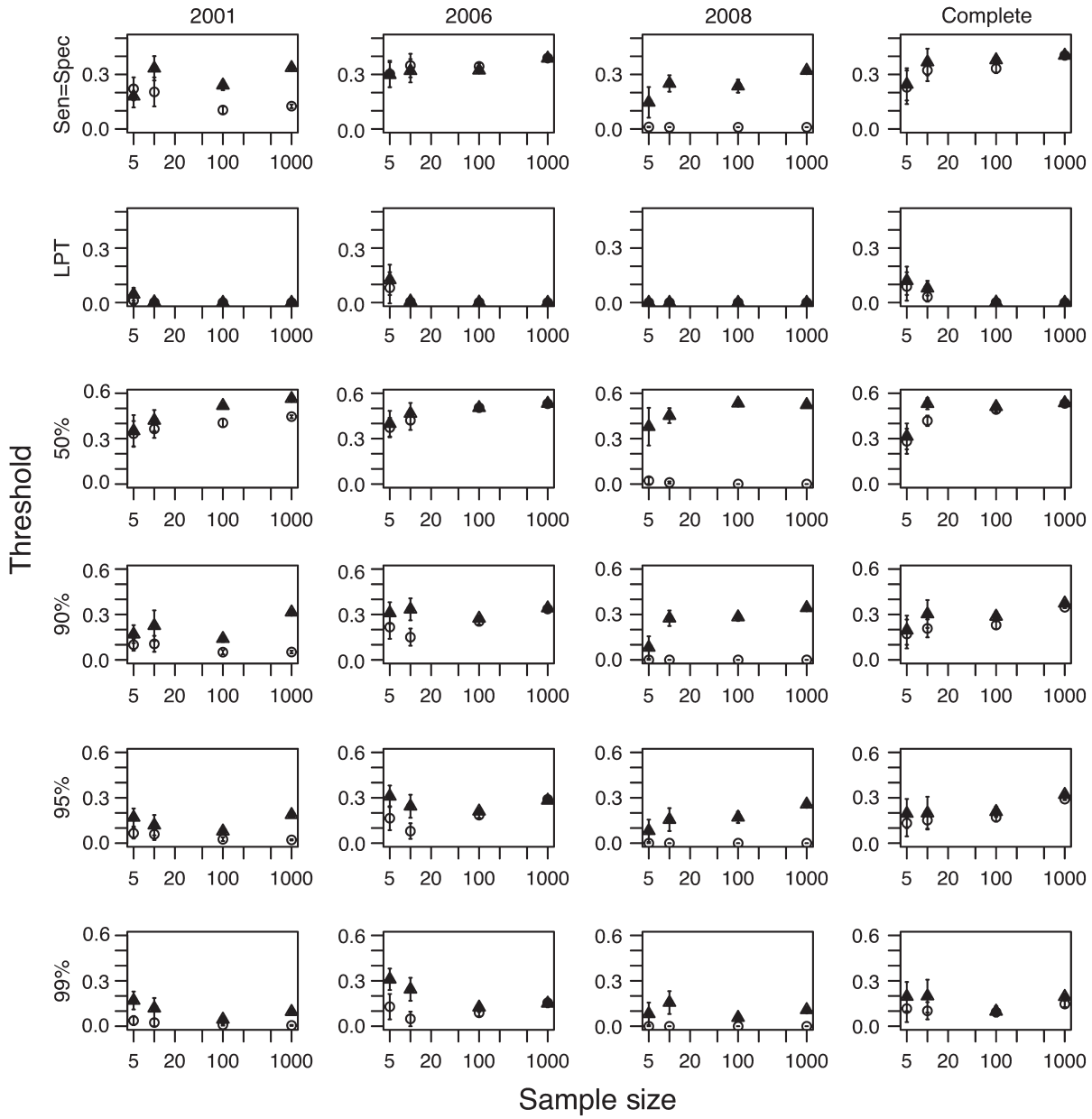

Figure 3. Threshold calculated for each model with a variety of metrics. Thresholds were selected by setting the sensitivity equal to specificity (method 1), setting the threshold to the lowest probability measured in the species occurrence data (method 7), and setting a threshold to a required level of sensitivity (method 8, sensitivities of 50, 90, 95, 99%). Additional threshold selection criteria results are shown in Supplementary material Appendix 1. Triangles represent threshold selected based on limited data. Circles represent thresholds selected from the multi-year distribution. In biased data (2001, 2006, 2008), thresholds required for a level of sensitivity were always lower than thresholds estimated based on limited data. In 2008, even achieving 50% sensitivity was impossible, but thresholds calculated based on limited data suggested relatively high thresholds. Error bars are + /– two standard errors across the 10 iterations for each sample size and year.

the distribution models. Liu et al. (2005), using complete knowledge of the distribution of two European species of plants, showed that setting the threshold to the ratio of presence points to total points in the original model performed best across multiple accuracy metrics and prevalence. They endorsed the specificity-sensitivity methods used in this study as well. Jimenez-Valverde and Lobo (2007) assessed model response to different threshold selection techniques across a range of sample sizes. They, too, recommended setting the threshold to the modeled data prevalence. This study confirmed that these are the best approaches for threshold selection. With most Maxent models – where prevalence tends to be quite low – the various sensitivity-specificity methods perform comparably to each other, as do those that maximize accuracy, and all of these perform better than the LPT method.

Hernandez et al. (2006) used the maximum of the sum of specificity and sensitivity to select a threshold and evaluate model performance under a range of sample sizes. They showed that Maxent could be effective with as few as 25 data points. In all three studies, thresholds were selected and accuracy assessed using the best data available. Our study shows that, while these approaches to threshold selection are better than others available, and each performs similarly

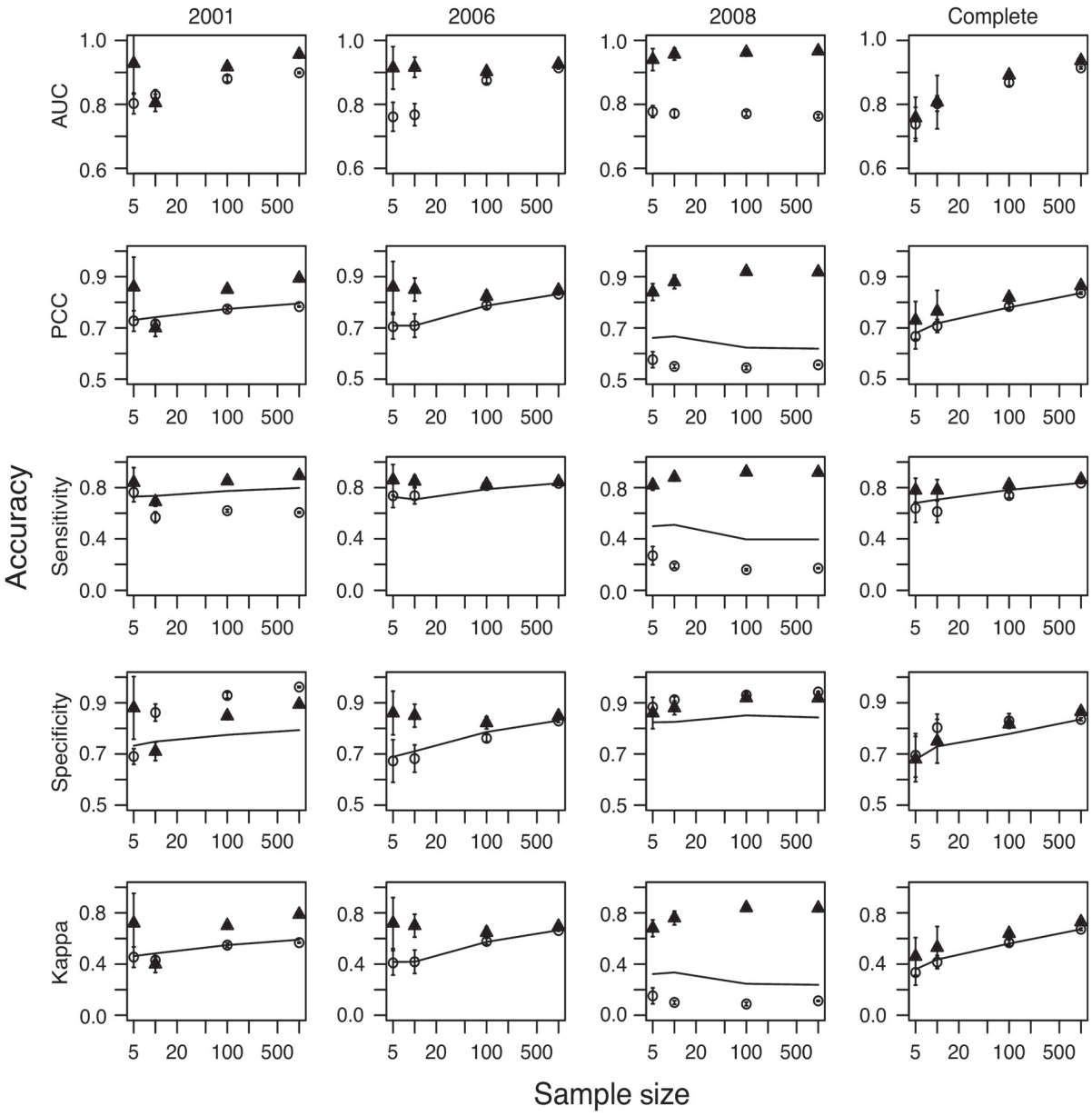

Figure 4. Accuracy calculated for each model. Threshold in each case was calculated by setting sensitivity equal to specificity (for threshold calculated from other methods, see Supplementary material Appendix 1). Triangles show accuracy calculated from limited data – how accurate the model appeared to be. Circles represent true accuracy, calculated from multi-year data – how accurate the model actually was. Lines represent accuracy if threshold had been calculated from multi-year data – how accurate the model could be. AUC lacks the line because it is a threshold-independent measure of accuracy. 2008 shows the greatest difference between reported and actual accuracy, whereas the multi-year data (i.e. unbiased) set shows little difference between the two. Error bars are + /– two standard errors for 10 iterations of each sample size and year.

well, an optimal threshold is more difficult to discern with limited data. Further, accuracy reported from limited data is frequently inflated above the true accuracy of the model. This was true both for the threshold-dependent accuracy metrics, and for AUC. Such problems are consistent with recent work showing that AUC, while independent of a threshold, suffers from problems similar to other accuracy metrics (Lobo et al. 2008).

While sample size has been explored thoroughly for most modeling approaches, spatial or temporal sample bias in sampling is a less-studied problem in species distribution models. Sampling bias is often caused by researcher access, and may have different implications for modeling outcome than temporal bias in a species not at equilibrium. That is, a species undergoing range contraction, or at a relative population minimum, may be found only in areas of high habitat quality. In such a case, a distribution model may be unnecessarily restrictive, although it may adequately represent the relative values of habitat. Sample bias also creates restricted distribution models (Hortal et al. 2008). In the case of sample bias, the restricted distribution will not necessarily correspond with high quality habitat for the species.

The effects of sampling bias may be as or more important than those observed for small sample sizes (Phillips et al. 2009). Indeed, models in this study that suffered from sample bias either in time (2001) or in space (2008) showed greater differences between the estimated and true model accuracy. Further, they also showed greater differences in calculated and optimal thresholds. Models created from a bias in time (i.e. a species at a population low) were overall more accurate than models created from researcher bias. On the other hand, models created from data with low sample bias (i.e. evenly sampled across 2006 and the multi-year distribution) showed that, even with small sample sizes, there was little difference between the threshold calculated from limited data and the optimal threshold. The only significant difference was at the lowest sample sizes, where accuracy was consistently inflated. In these cases, where data has been sampled evenly but only a few samples are available, it may be safe to say that, while the models may not be as accurate as they are reported to be, they are as accurate as they could be given the limitations of the model (i.e. despite the small sample size, the threshold was successfully calculated for optimal accuracy). In other words, well-sampled data with few records are better than biased data of any sample size. In fact, biased data appeared to perform better at low sample sizes. With small samples, there was a less tight model fit, which allowed for greater probabilities outside of the sampling area. Large samples, on the other hand, constrained the model to the sampling area, creating a worse overall map of distribution.

We believe our study sheds new light on challenges created when small and biased datasets are used for modeling species distributions, but several questions remain. The ratio of appropriate model-training data to validation data is critical, dependent on the application (Fielding and Bell 1997). Given a limited sample of species occurrences, how many points should be used in developing the model, and how many points should be reserved to select a threshold and calculate its accuracy? Our results suggest that model accuracy and the difference in reported and true accuracy tend to be related. As sample size increases, model accuracy increases, and the difference between reported and true accuracy decreases. Further, in cases where data is unbiased, the selected threshold differs little from the optimal threshold. In general, we suggest it is better to create a more accurate model with an inflated accuracy calculation than a less accurate model with a more accurate accuracy statistic: we would rather not know how good the good model is than know for certain how bad the bad model is. Regardless, more recent methods (e.g. jack-knife) may be a better approach for complete model evaluation with small samples. Of course, the type of model accuracy desired depends almost entirely on the application. False absences modeled into conservation prioritizations could be catastrophic, although resources spent on false presences may be just as problematic (Loiselle et al. 2003). In any research, we reiterate that it is critical to state the purpose and method for selecting a threshold, and to explore the sensitivity of the model to the threshold selected.

Acknowledgements – We thank S. Butterfield and The Nature Conservancy for funding and logistical support; L. Saslaw and J. Hurl and the Bureau of Land Management for logistical support; USDA-AFRI and the Dept of Environmental Science, Policy, and Management at the Univ. of California – Berkeley for funding; S. Beissinger, M. Kelly, L. Prugh, E. Rubidge, A. C. Burton, C. Golden, S. Sawyer, and W. Linklater for helpful comments and advice. Finally, we are grateful to three anonymous reviewers for their insightful comments.

References

Beyer, H. L. 2004. Hawths analysis tools for ArcGIS. – <www.spatialecology.com>.

Braun, S. E. 1985. Home range and activity patterns of the giant kangaroo rat, *Dipodomys ingens*. – J. Mammal. 66: 1–12.

Brown, K. A. et al. 2008. Multi-scale analysis of species introductions: combining landscape and demographic models to improve management decisions about non-native species. – J. Appl. Ecol. 45: 1639–1648.

Cantor, S. B. et al. 1999. A comparison of C/B ratios from studies using receiver operating characteristic curve analysis. – J. Clin. Epidemiol. 52: 885–892.

Cohen, J. 1960. A coefficient of agreement for nominal scales. – Educ. Psychol. Meas. 20: 37–46.

Cramer, J. S. 2003. Logit models from economics and other fields. – Cambridge Univ. Press.

De Marco, P. et al. 2008. Spatial analysis improves species distribution modelling during range expansion. – Biol. Lett. 4: 577–580.

DeMatteo, K. E. and Loiselle, B. A. 2008. New data on the status and distribution of the bush dog (*Speothos venaticus*): evaluating its quality of protection and directing research efforts. – Biol. Conserv. 141: 2494–2505.

Elith, J. and Leathwick, J. R. 2009. Species distribution models: ecological explanation and prediction across space and time. – Annu. Rev. Ecol. Evol. Syst. 40: 677–697.

Elith, J. et al. 2006. Novel methods improve prediction of species distributions from occurrence data. – Ecography 29: 129–151.

ESRI 2008. ArcGIS 9.1. – ESRI, Redlands, CA.

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

Freeman, E. A. and Moisen, G. G. 2008a. A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. – Ecol. Model. 217: 48–58.

Freeman, E. A. and Moisen, G. 2008b. PresenceAbsence: an R package for presence absence analysis. – J. Stat. Softw. 23: 31.

Gibson, L. et al. 2007. Dealing with uncertain absences in habitat modelling: a case study of a rare ground-dwelling parrot. – Divers. Distrib. 13: 704–713.

Graham, C. H. and Hijmans, R. J. 2006. A comparison of methods for mapping species ranges and species richness. – Global Ecol. Biogeogr. 15: 578–587.

Graham, C. H. et al. 2004. Integrating phylogenetics and environmental niche models to explore speciation mechanisms in dendrobatid frogs. – Evolution 58: 1781–1793.

Guisan, A. and Theurillat, J. P. 2000. Equilibrium modeling of alpine plant distribution: how far can we go? – Phytocoenologia 30: 353–384.

Guisan, A. et al. 1998. Predicting the potential distribution of plant species in an Alpine environment. – J. Veg. Sci. 9: 65–74.

Hanley, J. A. and McNeil, B. J. 1982. The meaning and use of the area under a receiver operating characterisic (ROC) curve. – Radiology 143: 29–36.

Hernandez, P. A. et al. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. – Ecography 29: 773–785.

Hortal, J. et al. 2008. Historical bias in biodiversity inventories affects the observed environmental niche of the species. – Oikos 117: 847–858.
Huntley, B. et al. 1995. Modelling present and potential future ranges of some European higher plants using climate response surfaces. – J. Biogeogr. 22: 967–1001.
Jimenez-Valverde, A. and Lobo, J. M. 2007. Threshold criteria for conversion of probability of species presence to either-or presence-absence. – Acta Oecol. 31: 361–369.
Kremen, C. et al. 2008. Aligning conservation priorities across taxa in Madagascar with high-resolution planning tools. – Science 320: 222–226.
Liu, C. R. etal. 2005. Selecting thresholds of occurrence in the prediction of species distributions. – Ecography 28: 385–393.
Lobo, J. M. et al. 2008. AUC: a misleading measure of the performance of predictive distribution models. – Global Ecol. Biogeogr. 17: 145–151.
Loiselle, B. A. et al. 2003. Avoiding pitfalls of using species distribution models in conservation planning. – Conserv. Biol. 17: 1591–1600.
Manel, S. et al. 1999. Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. – Ecol. Model. 120: 337–347.
Manel, S. et al. 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. – J. Appl. Ecol. 38: 921–931.
Papes, M. and Gaubert, P. 2007. Modelling ecological niches from low numbers of occurrences: assessment of the conservation status of poorly known viverrids (Mammalia, Carnivora) across two continents. – Divers. Distrib. 13: 890–902.
Pearson, R. G. et al. 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. – J. Biogeogr. 34: 102–117.
Phillips, S. J. et al. 2006. Maximum entropy modeling of species geographic distributions. – Ecol. Model. 190: 231–259.

Supplementary material (Appendix E6545 at <www.oikosoffice.lu.se/appendix>). Appendix 1.

Phillips, S. J. et al. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. – Ecol. Appl. 19: 181–197.
Raxworthy, C. J. et al. 2007. Applications of ecological niche modeling for species delimitation: a review and empirical evaluation using day geckos (*Phelsuma*) from Madagascar. – Syst. Biol. 56: 907–923.
Thorn, J. S. et al. 2009. Ecological niche modelling as a technique for assessing threats and setting conservation priorities for Asian slow lorises (Primates: Nycticebus). – Divers. Distrib. 15: 289–298.
Thuiller, W. 2004. Patterns and uncertainties of species' range shifts under climate change. – Global Change Biol. 10: 2020–2027.
USDA 2008. Soil Survey Geographic (SSURGO) Database for Carrizo Plain, CA. – <http://soildatamart.nrcs.usda.gov>, accessed 1 May 2008.
USDA Forest Service 2008. Existing vegetation – CALVEG. – USDA Forest Service, Region 5 Remote Sensing Lab, Mc Clelland, CA.
USGS 1994. Precipitation isohyets of California. – <http://dot.ca.gov>, accessed 1 May 2008.
USGS 2008. Shuttle Radar Topography Mission, 1 Arc Second scene. – <http://srtm.usgs.gov/>, accessed 1 May 2008.
Waltari, E. and Guralnick, R. P. 2009. Ecological niche modelling of montane mammals in the Great Basin, North America: examining past and present connectivity of species across basins and ranges. – J. Biogeogr. 36: 148–161.
Ward, D. F. 2007. Modelling the potential geographic distribution of invasive ant species in New Zealand. – Biol. Invasions 9: 723–735.
Williams, D. F. 1992. Geographic distribution and population status of the giant kangaroo rat, *Dipodomys ingens* an endangered and sensitive species of the San Joaquin Valley, California. – California Energy Commission.
Wisz, M. S. and Guisan, A. 2009. Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data. – BMC Ecol. 9: 8.