ECOGRAPHY

Research

The challenge of modeling niches and distributions for data-poor species: a comprehensive approach to model complexity

Peter J. Galante, Babatunde Alade, Robert Muscarella, Sharon A. Jansa, Steven M. Goodman and Robert P. Anderson

P. J. Galante (http://orcid.org/0000-0002-7025-3551) (pgalante@amnh.org), Babatunde A. and R. P. Anderson, Dept of Biology, City College of New York, City Univ. of New York, New York, NY, USA. Present address of PJG: Center for Biodiversity and Conservation, American Museum of Natural History, New York, NY, USA. RPA also at: Program in Biology, Graduate Center, City Univ. of New York, New York, NY, USA, and Division of Vertebrate Zoology (Mammalogy), American Museum of Natural History, New York, NY, USA. – R. Muscarella, Section for Ecoinformatics and Biodiversity, Dept of Bioscience, Aarhus Univ., Aarhus, Denmark. – S. A. Jansa, Bell Museum of Natural History and Dept of Ecology, Evolution, and Behavior, Univ. of Minnesota, St Paul, MN, USA. – S. M. Goodman, Field Museum of Natural History, Chicago, IL, USA, and Association Vahatra, Madagascar.

Ecography
41: 726–736, 2018
doi: 10.1111/ecog.02909

Subject Editor: Timothy Keitt
Editor-in-Chief: Miguel Araújo
Accepted 5 June 2017

Models of species ecological niches and geographic distributions now represent a widely used tool in ecology, evolution, and biogeography. However, the very common situation of species with few available occurrence localities presents major challenges for such modeling techniques, in particular regarding model complexity and evaluation. Here, we summarize the state of the field regarding these issues and provide a worked example using the technique Maxent for a small mammal endemic to Madagascar (the nesomyine rodent *Eliurus majori*). Two relevant model-selection approaches exist in the literature (information criteria, specifically AICc; and performance predicting withheld data, via a jackknife), but AICc is not strictly applicable to machine-learning algorithms like Maxent. We compare models chosen under each selection approach with those corresponding to Maxent default settings, both with and without spatial filtering of occurrence records to reduce the effects of sampling bias. Both selection approaches chose simpler models than those made using default settings. Furthermore, the approaches converged on a similar answer when sampling bias was taken into account, but differed markedly with the unfiltered occurrence data. Specifically, for that dataset, the models selected by AICc had substantially fewer parameters than those identified by performance on withheld data. Based on our knowledge of the study species, models chosen under both AICc and withheld-data-selection showed higher ecological plausibility when combined with spatial filtering. The results for this species intimate that AICc may consistently select models with fewer parameters and be more robust to sampling bias. To test these hypotheses and reach general conclusions, comprehensive research should be undertaken with a wide variety of real and simulated species. Meanwhile, we recommend that researchers assess the critical yet underappreciated issue of model complexity both via information criteria and performance on withheld data, comparing the results between the two approaches and taking into account ecological plausibility.

NORDIC SOCIETY OIKOS

www.ecography.org

Introduction

Ecological niche models (ENMs; often termed species distribution models, SDMs) constitute an important tool in ecology and evolution, but their application is especially difficult for species that have only few occurrences available. Correlative ENMs determine the relationship between localities where a species is known to occur, and the abiotic (e.g. climatic, edaphic) and biotic (e.g. vegetation, species interactions) properties of these locations, yielding an estimate of suitability for the species (Elith et al. 2006). These models have seen wide application throughout environmental biology, including use in conservation, agriculture, zoonotic diseases, and many other areas (Elith and Leathwick 2009). When copious data are available, ENMs even can be used for studying the variables driving population abundance, demography, and recruitment (Williams et al. 2009, Martínez-Meyer et al. 2013, Searcy et al. 2015, Muscarella and Uriarte 2016). Linked with such demographic information as well as genetic data, ENMs also hold promise for studying the effects of past and ongoing climate change (Pearson and Dawson 2003, Waltari et al. 2007).

Unfortunately, however, the majority of species are known from very few occurrences (Soberón et al. 2000), greatly hindering the production of useful ENMs for such species. Even in these cases, researchers often hope to use ENMs for various uses. Critically, ENM predictions inform conservation assessments, which are especially important for data-poor species (Anderson and Martínez-Meyer 2004, Franklin 2010). For these species, such models can help target geographic areas for future sampling. Similarly, newly detected invasive species or taxa associated with emerging zoonotic diseases present opportunities for contributions via ENMs, yet often afford only few occurrences. More generally, including data-poor species is necessary for comprehensive analyses identifying areas for preservation (i.e. reserve selection; Papeş and Gaubert 2007, Lawler et al. 2011), understanding macroecological patterns, and synthetic studies of the effect of past environmental change on patterns of phyloendemism (Brown et al. 2016, Prates et al. 2016).

Model complexity

Estimating optimal levels of model complexity remains a key outstanding methodological issue for ENMs, and especially challenging for species with few occurrences (Merow et al. 2013, Muscarella et al. 2014, Warren et al. 2014, Moreno-Amat et al. 2015). Model complexity has key relevance both for geographic predictions and identification of realistic hypotheses regarding the driving environmental variables (Araújo and Guisan 2006, Elith and Graham 2009, Merow et al. 2014). Major complications identifying optimal complexity exist even for the commonly used machine-learning algorithm Maxent (Anderson and Gonzalez 2011, Shcheglovitova and Anderson 2013), which has shown high performance for small sizes compared with other techniques (Hernandez et al. 2006, Wisz et al. 2008). Specifically, the Maxent software allows for the use of default settings for factors that affect model complexity greatly (e.g. feature classes and regularization multipliers; Phillips and Dudík 2008). In contrast, users can create models with a wide range of settings for each species (yielding many candidate models) to identify those that lead to optimal levels of model complexity. Although rarely done (Halvorsen 2013), such species-specific tuning of model settings (also termed 'smoothing'; Elith et al. 2010) has been shown to result in simpler and substantially more realistic Maxent models than those built using default settings (Anderson and Gonzalez 2011, Warren and Seifert 2011, Jueterbock et al. 2013, Muscarella et al. 2014, Radosavljevic and Anderson 2014, Warren et al. 2014, Moreno-Amat et al. 2015).

However, no general consensus yet exists regarding the most appropriate way to select optimal complexity for Maxent or many other ENMs (i.e. that best approximating the calibration data while holding the greatest generality when applied to independent data; Warren et al. 2008, Elith et al. 2011). Two main approaches involve evaluating model performance via internal testing (i.e. on calibration data), versus quantifying model performance on external (withheld) evaluation data. Regarding the first approach, some studies have advocated the use of information criteria, specifically Akaike's information criterion in the selection of optimally complex ENMs (AICc, corrected for small sample size; Baldwin 2009, Warren and Seifert 2011). However, although properly applied to several regression-based ENMs (e.g. GAM/GLM; Guisan and Thuiller 2005), information criteria do not fit the machine-learning paradigm perfectly (Warren and Seifert 2011). Specifically, the degrees of freedom for each model cannot be calculated exactly, and this issue is compounded because some feature classes can be penalized multiple times (Dudík 2007, Warren et al. 2014). Nevertheless, even if imperfect in this context, AICc still may be useful for machine-learning because it gives a quantitative measure without the use of external evaluation data, balancing model complexity with goodness-of-fit (Guisan and Thuiller 2005). In contrast, in the second approach, external performance is measured on withheld data, quantifying the model's ability to predict evaluation records. In this vein, two studies proposed estimating optimal complexity for Maxent models based on a sequence of two criteria for evaluation of withheld data (hereafter termed jackknife, for application with small sample sizes of occurrence records). In those particular implementations, the first criterion (e.g. omission rate) minimizes overfitting to calibration data, and the second one (e.g. AUC) maximizes discriminatory ability (Shcheglovitova and Anderson 2013, Radosavljevic and Anderson 2014). Here, we quantify overfitting as any evaluation omission rates higher than that expected based on the thresholding rule applied (see also Model selection techniques in Methods).

Present experiment and questions

Because progress in the field requires comparisons of these two approaches, we provide a worked example illustrating an integrated effort to compare these selection approaches. For the withheld-data approach, we use particular evaluation metrics in a sequential manner (i.e. first omission rate and then AUC; see below for details) but note that many other implementations are possible (Peterson et al. 2011, Warren and Seifert 2011). Although numerous studies have implemented either AICc or the jackknife method to Maxent model tuning, to our knowledge the two have not yet been directly compared.

Specifically, we address these issues with a poorly known forest-dwelling species endemic to wet montane areas of Madagascar, *Eliurus majori* (family Nesomyidae, subfamily Nesomyinae). We do so by creating a suite of candidate models built with a range of settings for factors that affect model complexity (feature classes and regularization multipliers). Out of that suite of candidate models, we compare those identified as optimal by each of the two approaches, as well as with the one produced using Maxent's default settings. For all three models, we also made qualitative assessments of their ecological plausibility based on our existing, although limited, knowledge of the study species (Goodman et al. 2014).

Additionally, we address one important possible confounding factor: the effects of sampling bias. Although both AICc and the jackknife method used on withheld data assume that occurrence data derive from unbiased sampling, such an assumption is likely violated in most datasets (Hijmans et al. 2000, Phillips et al. 2009, Peterson et al. 2011). Therefore, to assess whether either model-selection technique is affected strongly by spatial sampling bias (which likely results in environmental bias; Reddy and Dávalos 2003, Kadmon et al. 2004), we conduct these experiments with two datasets that should reflect different levels of sampling bias (Boria et al. 2014). Specifically, we use datasets of localities with and without application of a spatial filter. The original dataset, comprised of all localities, presumably reflects relatively high sampling bias in geography, typical of museum biodiversity data (Reddy and Dávalos 2003, Graham et al. 2004). To produce a second dataset, we spatially filter the localities, yielding a dataset that should reflect relatively less sampling bias (Veloz 2009, Anderson and Raza 2010, Carroll 2010, Hijmans 2012, Boria et al. 2014).

The present worked example addresses three main questions.

Question 1: how do model complexity and geographic predictions differ between default settings and each of the two model-selection techniques?

Maxent's default settings have a tendency to produce overfit models, leading to the following expectations. For each dataset (unfiltered or filtered), we expect that each model-selection technique (AICc or jackknife) will identify models that are simpler (fewer parameters) and less overfit (showing lower omission rates) than the one made using default settings. Similarly, we expect the geographic predictions of models selected using the two techniques to differ from those of the default models (low similarity, resulting in low Schoener's *D*-value and binary concordance; defined below).

Question 2: how do model complexity and geographic predictions differ between the two model-selection techniques (AICc and jackknife)?

We have no a priori expectation that the selection techniques will differ. Therefore, for each dataset, we expect that the two techniques will identify similar model complexity (number of parameters and measure of overfitting) and geographic predictions (high *D*-value and binary concordance).

Question 3: how do the results of these comparisons differ depending on whether or not the occurrence localities are spatially filtered?

For each selection technique, as well as the default settings, we expect that the datasets (unfiltered and filtered) will lead to different geographic predictions. Specifically, because the assumed bias is higher in the unfiltered dataset (likely resulting in increased complexity), the binary predictions for the unfiltered dataset should indicate a smaller area as suitable than for the filtered dataset.

Methods

Input data

We compiled locality information for *Eliurus majori* from museum voucher specimens (Supplementary material Appendix 1 Table A1). This largely arboreal species is endemic to Madagascar and only known from intact or fairly intact mesic montane forests of the eastern and central portions of the island (Soarimalala and Goodman 2011, Goodman et al. 2014). Identifications were inferred from a phylogeny based on the mitochondrial gene cytochrome b in which individuals assigned to *E. majori* were most closely related to specimens from that clade (with morphological confirmation by specialist S. M. Goodman) than to those from any other *Eliurus* species (Jansa et al. 1999, Jansa unpubl.; n = 23 unique localities). To reduce the likely effects of sampling bias in this dataset, we spatially filtered the 23 original localities such that the maximum number of localities was retained. Because of the heterogeneous vegetational and topographical landscape of Madagascar and the inferred likely level of sampling bias across geography (Goodman et al. 2014), we used a 10 km filtering distance, resulting in 13 occurrences (Boria et al. 2014). To do so, we used a preliminary version of spThin (functionally equivalent to ver. 0.1; Aiello-Lammens et al. 2015) in R (R Development Core Team) to sample the unfiltered dataset 10 000 times, and then randomly select one of the datasets that produced the maximum number of occurrence localities remaining (n = 13). These analyses do

not allow for tests of expectations regarding the level of complexity or overfitting between unfiltered and filtered datasets. Such tests would require sample-size rarefaction experiments and spatially independent evaluations (Boria et al. 2014). Rather, to address possible sensitivity of the model-selection techniques to biased sampling, we conducted all analyses first with the unfiltered dataset and then with the spatially filtered one.

As a set of plausible predictors likely to affect the species' distributions, we used 19 bioclimatic variables from WorldClim.org (Hijmans et al. 2005) at 30 arc seconds resolution. The WorldClim data describe aspects of temperature and precipitation, and have been shown to produce informative niche models of non-volant mammals (Jezkova et al. 2009, Anderson and Raza 2010). They are likely relevant for modeling this species, which appears to be associated with mesic montane conditions (Goodman et al. 2014). Clearly, other types of variables (e.g., soils, topography, vegetation, land use, and even biotic interactions) often constitute important additional determinants of species distributions (Elith and Leathwick 2009, Meier et al. 2010). However, using solely climatic variables (which hold pervasive effects on species distributions; Araújo et al. 2009) likely can provide an informative first-pass in estimating environmental constraints affecting geographic distributions, especially when little is known of the biology of a rare species. To the degree that other factors not considered here affect the species' distribution, the models would be underspecified (see discussion later regarding interpretation of underfit and overfit models based on quantitative evaluations). Note that even though 19 variables were input here, not all of them were necessarily used for any feature class, and some of them might be used repeatedly for hinge features (Phillips and Dudík 2008). We restricted the selection of environmental data from 'background' pixels to a region in which known records are more likely to form a representative sample of the climatic conditions suitable for the species (Anderson and Raza 2010, Peterson et al. 2011, Anderson 2013). Specifically, we used a rectangle encompassing a 100 km buffer around the most extreme locality in each of the four cardinal directions. Many methods of selecting the background exist in the literature (Phillips 2008, Anderson and Raza 2010, Peterson et al. 2011, Radosavljevic and Anderson 2014), but the bounding rectangle we use achieves our primary aim of excluding areas that are climatically suitable, but to which *E. majori* has been unable to disperse and/or is not known to occur.

Niche modeling

We created niche models allowing for a wide range of complexity by varying two critical settings: feature classes (FCs) and regularization multiplier (RM). The various FCs allowed in a given Maxent model control the flexibility of the shape of the modeled response to each input variable. Complementarily, regularization promotes simplicity by applying penalties for additional parameters included in a model, and higher weights for them (Phillips and Dudík 2008, Phillips et al. 2009, Merow et al. 2013). Hence, higher regularization protects against overfitting (the omission rate at which 10% (or at least one) of the localities are omitted). In particular, we created a suite of models by allowing increasing complexity of the FCs employed, as likely to be appropriate for the small sample size available for this species: linear (L); linear and quadratic (LQ); hinge (H); and linear, quadratic, and hinge (LQH). For each FC combination, we built models across a wide range of levels of regularization. By default, Maxent assigns a particular β regularization value for each feature class (Phillips 2008, Elith et al. 2011). Current releases of Maxent allow the use of a regularization multiplier, a single coefficient multiplied to each respective β value to increase or decrease the penalties assigned, across all feature classes in concert. Increasing penalties to Maxent models generally decreases the levels of overfitting, but excessive regularization will lead to underfitting. This corresponds to a decrease in the explanatory ability of the model (e.g. lower discrimination) and an increase in the likelihood of commission errors. Assessing a range of regularization multipliers allows for a wide breadth of complexity tests (Merow et al. 2013), and permits Maxent to perform within a varying range of complexities (Radosavljevic and Anderson 2014; see also results) that can lead to a best-fit model. Because our evaluation metrics aim to balance simplicity with explanatory ability, for each FC combination, we built models across a set of RM values that ranged from 1–5, increasing by increments of 0.25. As in recent studies using similar datasets, preliminary analyses (not shown) indicated that optimality was found within this range (Radosavljevic and Anderson 2014). This resulted in a suite of 68 combinations of FC/RM settings, yielding 68 candidate models.

We made the models in Maxent (ver. 3.3.3k) using the R package 'ENMeval' (ver. 0.1.1; Muscarella et al. 2014). We extracted AICc values, parameters used (lambda values; see below), a measure of overfitting (10% omission rate), and AUC (see below for details of evaluation statistics). To generate evaluation statistics based on withheld data, we employed the jackknife method of ENMeval because of the small sample size used in this study. We ran all models with a single set of 10 000 background pixels and chose the raw output format for all analyses (except visualization or comparisons in geographic space; see below). Note that neither omission rates (hereafter, OR) nor the area under the receiver operator characteristic curve (evaluation AUC; hereafter, AUC) values used in the sequential method (see below) differ among the various Maxent output formats, all of which preserve relative ranks. We did not explicitly allow Maxent to include as background pixels all of those where occurrence localities lay, using the 'noaddsamplestobackground' argument of Maxent (however, random background pixels may be sampled in these locations).

Additionally, to quantify concordance of resulting predictions in geographic space across the entirety of Madagascar (projecting well beyond the calibration study region; see below) we re-calibrated models in the Maxent graphic user interface. Here, we used all localities from each dataset (filtered and unfiltered; no withheld data) and either the default, AICc-optimal, or sequential-optimal settings (using the species-specific background region, and then projecting to the whole island). To allow comparisons among the three predictions, models were projected using the logistic format (see Royle et al. 2012, Hastie and Fithian 2013 for assumptions). Several methods exist to compare model outputs (e.g. Schoener's *D*, Moran's *I*, Spearman's rank correlation coefficient, and percent binary concordance; Schoener 1968, Warren et al. 2008). Among these, we first chose one parametric and one nonparametric measure for pairwise comparisons of the resulting predictions. First, we calculated Schoener's *D*-values characterizing the similarity in geographic space considering the entire gradient of prediction strength (using the R package 'dismo'; Hijmans et al. 2013). Values of *D* range from 0 to 1, with higher values indicating increased geographic concordance between predictions. Subsequently, to quantify the similarity of predicted suitable vs. unsuitable regions, we converted predictions to binary maps according to the 10% training omission-rate threshold of that model using the R package 'biomod2' (Thuiller et al. 2009). Many thresholding rules are justified for presence-only occurrence data (Peterson et al. 2011), and here we used one likely to be reasonable (10% calibration omission rate; see below). Using these binary maps, we calculated the proportion of pixels predicted present and measured binary concordance among predictions ('raster', R; Hijmans and van Etten 2012). Additionally, we measured the altitudinal range of each prediction to compare with elevations from specimen tags. Although Schoener's *D*-value and measures of binary concordance both range between 0 and 1, these statistics are not directly comparable in absolute terms. Rather, relative patterns must be interpreted within each metric separately. Hence, for each metric, the researcher must establish a study-region specific qualitative determination of the level of similarity interpreted as 'similar' and 'different'.

Model selection techniques

AICc technique

We first identified the optimal model based on AICc (corrected for small sample sizes; Warren and Seifert 2011), which scores models based on balancing complexity and goodness of fit. AICc penalizes high model complexity, giving the lowest (best) score to the model that best approximates the calibration data without being overly complex. Specifically, it measures complexity by the number of parameters actually included in each resulting model. The lambda coefficients of a Maxent model indicate weights for all included parameters (i.e. those with non-zero values; Phillips et al. 2006). Importantly, AICc leads to various related quantitative measures, including the change in AICc score (ΔAICc; the difference between the likelihood of a given model and that of the best model). Additionally, to allow comparisons of overfitting with the models selected via sequential criteria, for model settings selected as optimal by AICc, we obtained the average omission rate using the jackknife procedure and the same thresholding rule as for sequential criteria. AICc (rather than BIC) was used here because the available parameter space is massive, and we do not expect that any of the models' approximations of the data to be correct, only that one of the candidate models will have the least predictive error among those examined (Aho et al. 2014).

Jackknife with sequential criteria

Next, we identified the optimal model using a sequence of two criteria based on performance on withheld data. To obtain evaluation statistics, we partitioned localities using a jackknife technique, which is useful for small samples size of localities (small n). The jackknife (a method of withholding data) consists of n iterations; in each iteration n – 1 localities are used for calibration, and the model is evaluated on the withheld locality (Pearson 2007, Shcheglovitova and Anderson 2013). This was done for each combination of settings, with performance averaged across all n iterations for each measure of performance.

As sequential criteria for model selection, we first identified the models that displayed the lowest average OR and then, of that subset of models, we chose the one with the highest average AUC score (Shcheglovitova and Anderson 2013). Omission rates (the proportion of evaluation localities with predicted values below a particular threshold, e.g. 'omitted') indicate whether a model is overfit to the calibration data. Specifically, overfit models typically have higher than expected ORs. We calculated the omission rate on the (withheld) evaluation locality of each iteration after applying the 10% calibration omission rate threshold (and then averaged across jackknife iterations). For this thresholding rule, because approximately 10% of evaluation localities are expected to fall outside the resulting binary prediction, omission rates above 10% indicate overfitting (Pearson 2007, Shcheglovitova and Anderson 2013, Radosavljevic and Anderson 2014). Because of the presence-only nature of the occurrence records (and lacking tenable absence data), we were not able to calculate unbiased estimates of commission errors, or false predictions of presences. Notably, model selection based first on OR (to minimize omission rates) might tend to identify very permissive (potentially underfit) models with high commission errors. For this reason, we applied a somewhat relaxed thresholding rule (10% training omission, rather than the stricter minimum training presence rule). Additionally, the second criterion applied does give higher values to models that predict smaller areas (that should tend to have lower commission rates, even though these cannot be quantified here in an unbiased manner). Specifically, AUC (the area under the curve of the receiver operator characteristic plot) gives a relative measure of overall discriminatory ability, by quantifying the proportion of instances in which a

Table 1. Summary of optimal tuning experiments for Maxent models of the Malagasy rodent *Eliurus majori*. Results are provided for two model-selection techniques (AICc and jackknife) as well as the default settings, for two datasets (unfiltered and filtered localities). Settings, number of parameters (non-zero parameter values (lambda)), and the omission rate (OR) for evaluation localities is provided for all experiments.

	Unfiltered (n = 24)			Filtered (n = 13)		
	Settings	λ	10% OR	Settings	λ	10% OR
AICc	LQ1.5	6	0.1304	H3.5	2	0.1429
Jackknife	H1.5	14	0.1304	H4.25	3	0.1429
Default	LQH1	17	0.2609	LQ1	8	0.2143

randomly selected occurrence record ranks higher than a randomly selected background pixel (Peterson et al. 2011). AUC values were averaged across all jackknife iterations, yielding an average AUC for each combination of model settings. Overall, this sequential-selection method is designed to avoid models that are overfit to calibration data (via the OR selection criterion), but hence it only indirectly penalizes model complexity (in contrast to AICc).

Data deposition

Data available from the Dryad Digital Repository: <http://dx.doi.org/10.5061/dryad.t84q0> (Galante et al. 2017).

Results

Default settings versus selection techniques (Question 1)

As expected, for both datasets (spatially unfiltered and filtered) the optimal models selected by each selection technique had fewer parameters than models made with default settings (Table 1). There was a large difference in the number of parameters used between AICc-optimal and default models in each dataset; however, for both datasets, many variables used in the AICc-optimal model were also used in the default model. The same pattern was apparent for the jackknife technique (although weaker for the unfiltered dataset than the filtered one). For both datasets, all but one of the variables incorporated in the AICc-optimal model were also used in the default model. Similarly, all (unfiltered dataset) and all but one (filtered dataset) of the variables used in the jackknife-optimal model were also used in the default model (Supplementary material Appendix 1 Table A3). Likewise, optimal settings for each selection technique consistently led to models with lower overfitting than the models based on default settings. Whereas both of the selection techniques displayed evaluation ORs slightly above the theoretically expected 10% for this thresholding rule (Radosavljevic and Anderson 2014), those of the default models were far higher, indicating greater overfitting (Table 1, Fig. 1).

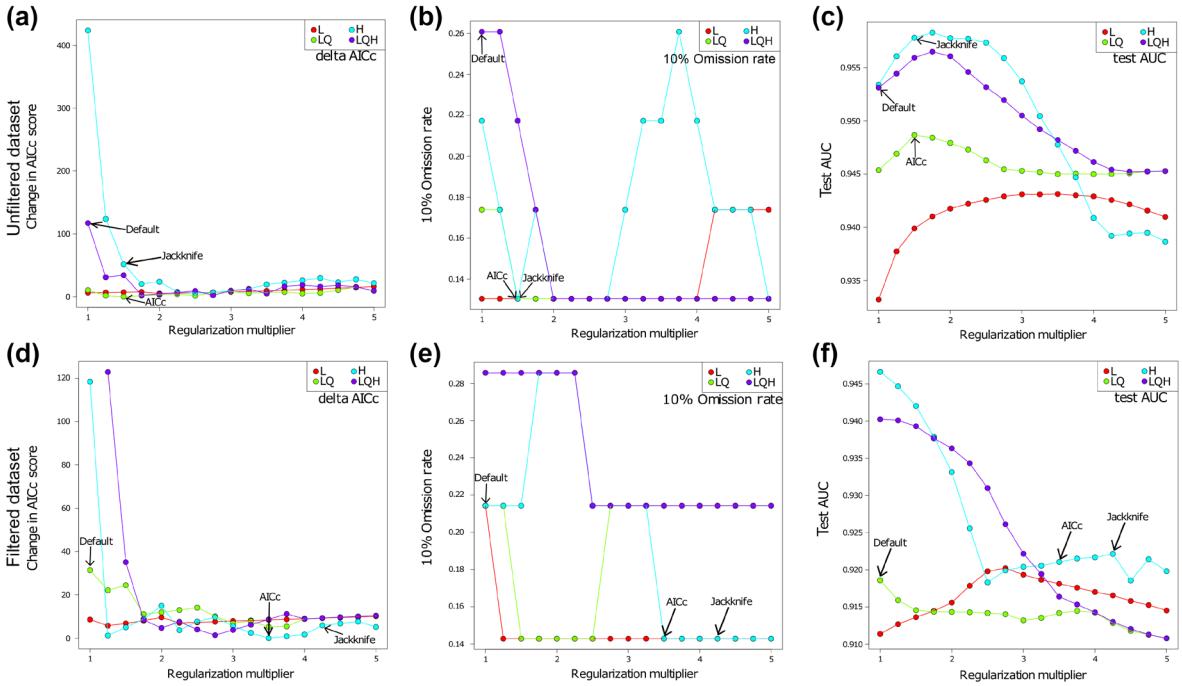

Figure 1. Three evaluation statistics for unfiltered (top row) and filtered (bottom row) locality datasets for *Eliurus majori* resulting from optimization of Maxent models. Left panels (a, f) show the difference in AICc scores between each model and the model that received the lowest AICc score. Middle panels (b, e) show omission rates of withheld data at the 10% calibration threshold. Right panels (c, f) show the test AUC values for each model. In each panel, arrows point to the optimal and default models, showing how changes in model settings can affect selection criteria. Model statistics are shown as feature classes (L = linear, LQ = linear + quadratic, H = hinge, LQH = linear + quadratic + hinge) increasing in regularization multiplier.

Regarding the geographic predictions, Schoener's *D*-values as well as binary concordance also matched expectations (Table 2). Specifically, the *D*-values indicated that the AICc-selected models and default models were quite different, for both datasets. Likewise, the corresponding comparisons of the jackknife technique and default models show low similarity, although somewhat higher using the unfiltered dataset. These trends in relative similarity for comparisons of continuous predictions were echoed in the corresponding comparisons of the binary maps.

Based on visual inspection, the two model-selection techniques showed marked differences from default models for one dataset, but not for the other. For the unfiltered dataset, the default model was noticeably, though not drastically different from the models selected by either technique, and all three models showed small, restrictive areas as suitable. These areas were mostly surrounding known occurrence localities. Both the model selected by the jackknife technique and that created using default settings predicted the same elevational range as suitable (454–2744 m), which was more restrictive than the range indicated by the model selected by AICc (171–2744 m). For the filtered dataset, models chosen by the two selection techniques were drastically different from that created using default settings. The default model was much more restricted to areas immediately around the occurrence localities, yet spanned a larger elevational range (171–2744 m), while the other two models showed much larger areal extent, yet were narrower in elevational range (442–2744 m).

AICc versus jackknife (Question 2)

The models selected as optimal using the two selection techniques had different levels of complexity and overfitting, as well as geographic predictions when using the unfiltered dataset, but these were similar using the filtered dataset (Table 1, Fig. 1). The unfiltered dataset led to a large discrepancy in the number of parameters used by optimal models, and many particular variables were not shared between optimal models. In contrast, the models selected using the filtered dataset incorporated very similar numbers of parameters. For the unfiltered dataset, only three of the six variables used by the AICc-optimal model were incorporated into the jackknife-optimal model. For the filtered dataset, both selection techniques used the same two variables (albeit for the jackknife-optimal model, Bio 8 was used twice as hinge features; Supplementary material Appendix 1 Table A3). With each of the respective datasets, the two selection techniques led to identical ORs. Matching the results regarding numbers of parameters, Schoener's *D*-values for the comparison between the models identified by each of the selection techniques was low for the unfiltered dataset but very high for the filtered dataset (Table 1). Again, the same pattern of relative values was found for the percent binary concordance, although the values were much higher.

The geographic predictions chosen by these two model-selection techniques were very similar (as judged by visual inspection) for one dataset, and less so for the other. Using the spatially unfiltered dataset, the model selected by the sequential technique was more broadly predictive, with stronger predictions in the higher elevations, and showed higher suitability in the mid-elevation areas. Using this dataset, the elevations predicted as present from a binary map of each prediction were also different. In contrast, using the spatially filtered dataset, models showed very high similarity in geographic space, such that the elevations predicted as present using binary maps of each prediction showed the same elevational range as suitable.

Table 2. Summary of comparisons of continuous (logistic output; A and B) and binary (C and D) model predictions in geographic space for tuning experiments for Maxent models of the Malagasy rodent *Eliurus majori*. Schoener's *D*-value (A and B) as well as binary concordance (C and D) are provided for all pairwise comparisons of two model-selection techniques (AICc and jackknife) as well as the default settings, for each of two datasets (spatially unfiltered and filtered localities).

	Jackknife	Default
Continuous comparisons (Schoener's *D*)		
(A) Unfiltered dataset		
AICc	0.817	0.792
Jackknife	–	0.852
(B) Filtered dataset		
AICc	0.941	0.749
Jackknife	–	0.710
Binary comparisons (thresholded predictions)		
(C) Unfiltered dataset		
AICc	96.2%	96.1%
Jackknife	–	98.7%
(D) Filtered dataset		
AICc	99.5%	95.7%
Jackknife	–	95.2%

Impact of spatial filtering (Question 3)

As expected, comparisons between datasets (spatially unfiltered vs filtered) showed marked differences in geographic predictions for each model-selection technique as well as for the default settings. The geographic agreement between continuous predictions was low, as quantified by Schoener's *D* (unfiltered compared with filtered: AICc = 0.6294; jackknife = 0.5846; default = 0.6827). As found above, binary concordance was consistently higher than the *D*-values in an absolute sense, but the values for these comparisons were low relative to the range of values found earlier for the comparisons between techniques for a given dataset (AICc = 95.10%; jackknife = 93.55%; default = 95.46%; Table 1). Furthermore, as predicted, models from the unfiltered dataset consistently indicated smaller areas as suitable than those from the filtered dataset, as quantified by the percent of Madagascar predicted suitable (AICc: 8.43% unfiltered, 10.92% filtered; jackknife: 5.04% unfiltered, 11.42% filtered; default: 4.71% unfiltered, 9.16% filtered). For the unfiltered dataset, both optimal models as well as the default one used predominantly temperature-related variables. For the filtered dataset, both optimal models as well as

the default model incorporated derived variables showing a fairly even mix of purely temperature variables with variables that combine temperature and precipitation information (e.g. mean temperature of the wettest quarter; Supplementary material Appendix 1 Table A3).

The models created in this experiment varied substantially in geographic space between spatially unfiltered and filtered datasets (Fig. 2). For the unfiltered dataset, the three models each predicted a small proportion of the island as highly suitable, with few differences among them. In those models, the particular areas with high suitability fell in eastern portions of the humid highlands. In contrast, use of the filtered dataset led to models that indicated a much larger proportion of the island as highly suitable, with clear differences existing among the three models. These differences were evident between the two datasets, within each selection technique. The maps from the two selection techniques (AICc and jackknife) displayed very high similarity. Each indicated strong prediction throughout humid areas of intermediate and high elevation (including the westernmost known locality for this species and other large extents of the highlands not strongly predicted in any of the models with the unfiltered dataset). Differing markedly from each of those two models, the one made using the default settings was largely restricted to the eastern humid highland regions, but not to the degree of any of the models made with unfiltered records.

Discussion

The comprehensive evaluation approach we illustrate here in the worked example for the Malagasy endemic rodent *E. majori* illuminated key trends for the two model-selection techniques. Following expectations, the two techniques produced simpler models (with fewer parameters) than did default settings. Furthermore, the lower omission rates obtained with AICc and the jackknife technique indicated lower levels of overfitting than when using default settings. We reiterate that, lacking true absence data, the present analyses cannot adequately quantify possible overpredictions. Nevertheless, these two model selection techniques also led to more realistic geographic predictions (including

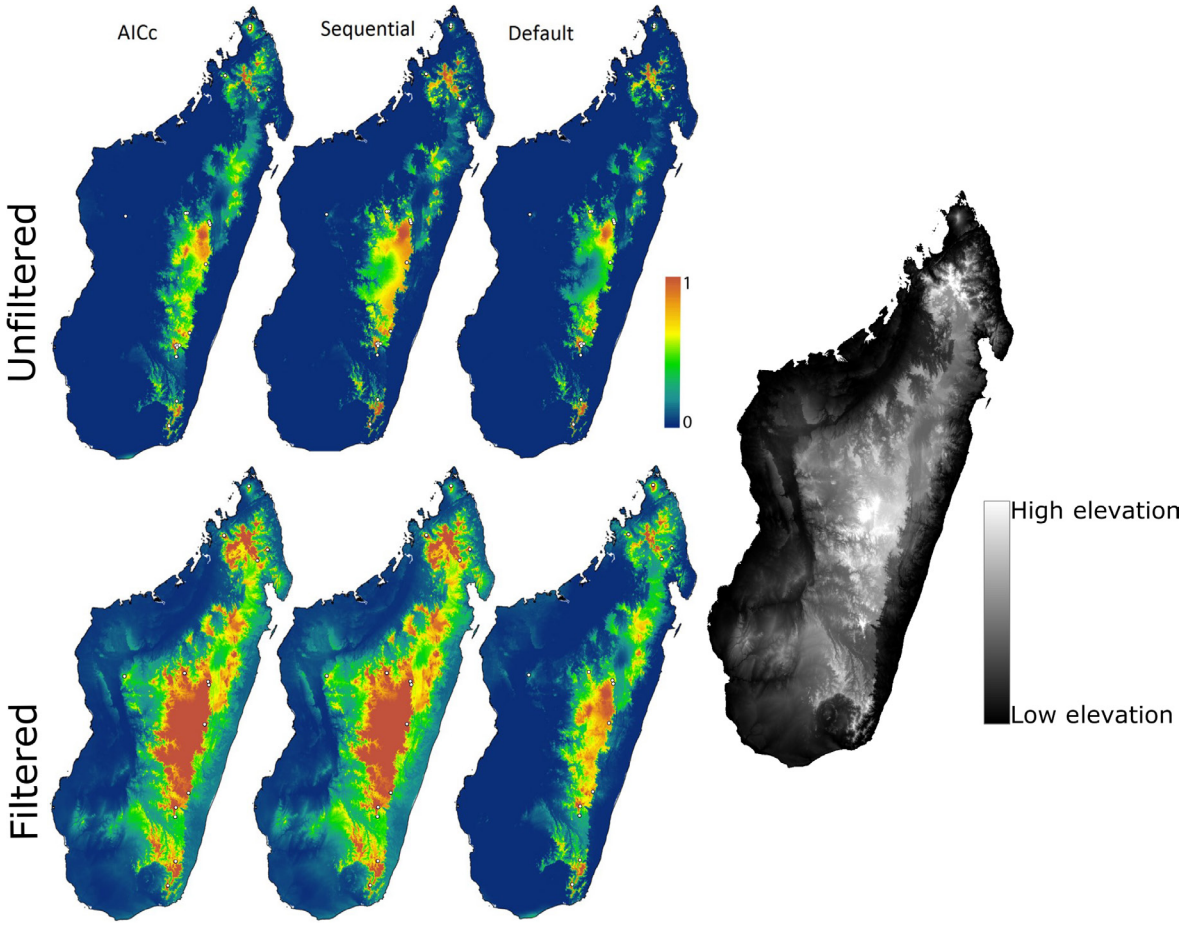

Figure 2. Optimal and default Maxent models for the Malagasy rodent *Eliurus majori* (logistic output). Results correspond to the unfiltered dataset (top) and filtered dataset (bottom), and three ways of determining model settings: AICc (left), sequential criteria based on performance on withheld (via jackknife) data (middle), and default settings (right). Plotted localities (white dots) represent unfiltered or filtered occurrence records for the corresponding row. Greyscale map shows elevation for reference (Hijmans et al. 2005).

consideration of possible under and over prediction as indicated by expert opinion). These trends were especially strong when using the spatially filtered dataset. As expected for both selection techniques, simpler models yielded larger suitable geographic areas when compared with default settings, particularly in the middle latitudes of the range of *E. majori*. Specifically, using the unfiltered dataset, both model selection techniques led to areas of higher prediction that were less concentrated around known records (likely due to sampling bias). The filtered dataset led to much more realistic geographic predictions for the highlands using both model selection techniques, in comparison with the default.

The above trends were strongest – and the selection techniques performed most similarly – for the filtered dataset (which is most likely to match the assumption of unbiased sampling). The model-selection techniques differed substantially regarding numbers of parameters for the unfiltered dataset (with AICc showing simpler models), but with the filtered dataset both techniques led to much simpler models than default. Regarding geographic predictions, none of the models created using the unfiltered dataset showed high suitability for the majority of the humid highlands, which includes the westernmost known locality of the species even though all models included multiple variables related to precipitation (Supplementary material Appendix 1 Table A3). These models all indicated that the major variable associated with environmental suitability (and hence, an identified putative driver) for this species is temperature (Supplementary material Appendix 1 Table A3). For the filtered dataset, in contrast, all models indicated a mix of variables solely reflecting temperature and those capturing interactions between temperature and precipitation. This reflects a more realistic assessment of the species based on our prior knowledge. The elevational ranges indicated by binary optimal models selected using this dataset were varied. The model selected by the jackknife technique indicated the same elevation range as the default model, both of which were more similar to the observed elevation range of this species (875–2500 m; Soarimalala and Goodman 2011) than was that of the AICc-selected model.

In contrast, for the filtered dataset, the predictions from both techniques were remarkably similar, with minor differences lying in the less suitable areas. For that dataset, the selection techniques both showed moderate suitability for nearly the entire highland area of the island, including the westernmost locality (Fig. 2). Reconstructions of vegetation types in Madagascar indicate that such regions held humid montane forests before extensive anthropogenic deforestation (Rakotondratsimba and Goodman 2014). Similarly, the elevational ranges indicated as optimal by both techniques using this dataset were both more similar to the observed elevations than was that for the default model. Hence, based on our prior knowledge of the study species, we determined that models made with the filtered dataset (especially those corresponding to the two selection techniques) more closely match reality than those made using the unfiltered dataset. However, these results also suggest that the behavior of these model-selection techniques depends on the level of sampling bias present, here, indicating that AICc may be more robust to departures from the assumption of unbiased sampling.

Future directions

These results show trends for a single data-poor species, and help set an agenda for future investigations that could lead to more general conclusions and recommendations. First, we advocate that researchers perform similar model-tuning experiments for the species at hand in a given study, also taking into account ecological plausibility (Franklin 2010). Second, the field needs strategic comprehensive studies aimed at discovering general patterns. This research should include similar experiments using: various types of relevant predictor variables (i.e. both climatic and non-climatic; abiotic and biotic), a range of sample sizes, other data-partitioning methods, varying levels of sampling bias, and modifications to the particular implementations of these selection techniques. For example, other implementations could include different thresholding rules, performance metrics (e.g. the difference between calibration and evaluation AUC values; Warren and Seifert 2011, Radosavljevic and Anderson 2014), or non-sequential weighting. Ideally, such experiments would include other real species from a variety of taxa, regions, and life history traits. Additionally, research using simulated species with known tolerances and varying levels of niche complexities may offer critical opportunities for addressing these issues. Together, studies using real data and those with simulated species – each across a wide range of sample sizes – would help elucidate more general conclusions regarding the performance of these two selection techniques. If possible, such conclusions would facilitate broad implementation of niche models for pressing uses across environmental biology (Peterson 2006, Anderson 2012). These types of investigations are crucial for studying general trends in biodiversity and conservation and are necessitated by the lack of occurrence data for the overwhelming majority of taxa (Soberón et al. 2000).

Acknowledgements – Robert A. Boria, Ana C. Carnaval, Maria Gavrutenko, Beth Gerstner, Michael J. Hickerson, Jamie M. Kass, and Mariano Soley-Guardia provided feedback and insight during the course of this project.
Funding – This research was made possible by funding from the National Science Foundation grants DEB-1119918 to SAJ, and DEB-1119915 to RPA, including a Research Experiences for Undergraduates supplement to support BA. BA obtained additional support from the City College Academy for Professional Preparation). RM was supported by NSF DBI-1401312.
Conflicts of interest – The authors declare no conflicts of interest in this submission.

References

Aho, K. et al. 2014. Model selection for ecologists: the worldviews of AIC and BIC. – Ecology 95: 631–636.

Aiello-Lammens, M. E. et al. 2015. spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. – Ecography 38: 541–545.

Anderson, R. P. 2012. Harnessing the world's biodiversity data: promise and peril in ecological niche modeling of species distributions. – Ann. N. Y. Acad. Sci. 1260: 66–80.

Anderson, R. P. 2013. A framework for using niche models to estimate impacts of climate change on species distributions. – Ann. N. Y. Acad. Sci. 1297: 8–28.

Anderson, R. P. and Martínez-Meyer, E. 2004. Modeling species' geographic distributions for preliminary conservation assessments: an implementation with the spiny pocked mice (*Heteromys*) of Ecuador. – Biol. Conserv. 116: 167–179.

Anderson, R. P. and Raza, A. 2010. The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus *Nephelomys*) in Venezuela. – J. Biogeogr. 37: 1378–1393.

Anderson, R. P. and Gonzalez, I. 2011. Species-specific tuning increases robustness to sampling bias in models of species distributions: an implementation with Maxent. – Ecol. Model. 222: 2796–2811.

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

Araújo, M. B. et al. 2009. Reopening the climate envelope reveals macroscale associants with climate in European birds. – Proc. Natl Acad. Sci. USA 106: E45–E46.

Baldwin, R. A. 2009. Use of maximum entropy modeling in wildlife research. – Entropy 11: 854–866.

Boria, R. A. et al. 2014. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. – Ecol. Model. 275: 73–77.

Brown, J. L. et al. 2016. Predicting the genetic consequences of future climate change: the power of coupling spatial demography, the coalescent, and historical landscape changes. – Am. J. Bot. 103: 1–11.

Carroll, C. 2010. Role of climatic niche models in focal-species-based conservation planning: assessing potential effects of climate change on northern spotted owl in the Pacific Northwest, USA. – Biol. Conserv. 143: 1432–1437.

Dudík, M. 2007. Maximum entropy density estimation and modeling geographic distributions of species. – Dissertation Abstracts Int. 68: 9.

Elith, J. and Graham, C. H. 2009. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. – Ecography 32: 66–77.

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.

Elith, J. et al. 2010. The art of modelling range-shifting species. – Methods Ecol. Evol. 1: 330–342.

Elith, J. et al. 2011. A statistical explanation of MaxEnt for ecologists. – Divers. Distrib. 17: 43–57.

Franklin, J. 2010. Mapping species distributions: spatial inference and prediction. – Cambridge Univ. Press.

Galante, P. J. et al. 2017. Data from: The challenge of modeling niches and distributions for data-poor species: a comprehensive approach to model complexity. – Dryad Digital Repository, <http://dx.doi.org/10.5061/dryad.t84q0>.

Goodman, S. M. et al. 2014. Small mammals or tenrecs (Tenrecidae) and rodents (Nesomyidae). – In: Goodman, S. M. and Raherilalao, M. J. (eds), Atlas of selected land vertebrates of Madagascar. Association Vahatra, Antananarivo, Madagascar, pp. 249–250.

Graham, C. H. et al. 2004. New developments in museum-based informatics and applications in biodiversity analysis. – Trends Ecol. Evol. 19: 497–503.

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

Halvorsen, R. 2013. A strict maximum likelihood explanation of MaxEnt, and some implications for distribution modelling. – Sommerfeltia 36: 1–132.

Hastie, T. and Fithian, W. 2013. Inference from presence-only data; the ongoing controversy. – Ecography 36: 864–867.

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.

Hijmans, R. J. 2012. Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. – Ecology 93: 679–688.

Hijmans, R. J. and van Etten, J. 2012. raster: geographic analysis and modeling with raster data. – R package ver. 2.0-12.

Hijmans, R. J. et al. 2000. Assessing the geographic representativenes of Genebank collections: the case of Bolivian wild potatoes. – Conserv. Biol. 14: 1755–1765.

Hijmans, R. J. et al. 2005. Very high resolution interpolated climate surfaces for global land areas. – Int. J. Climatol. 25: 1965–1978.

Hijmans, R. J. et al. 2013. dismo: species distribution modeling. – R package ver. 0.8-17.

Jansa, S. A. et al. 1999. Molecular phylogeny and biogeography of the native rodents of Madagascar (Muridae: Nesomyinae): a test of the single-origin hypothesis. – Cladistics 15: 253–270.

Jezkova, T. et al. 2009. Pleistocene impacts on the phylogeography of the desert pocket mouse (*Chaetodipus penicillatus*). – J. Mammal. 90: 306–320.

Jueterbock, A. et al. 2013. Climate change impact on seaweed meadow distribution in the North Atlantic rocky intertidal. – Ecol. Evol. 3: 1356–1373.

Kadmon, R. et al. 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. – Ecol. Appl. 14: 401–413.

Lawler, J. J. et al. 2011. Using species distribution models for conservation planning and ecological forecasting. – In: Drew, C. A. et al. (eds), Predictive species and habitat modeling in landscape ecology: concepts and applications. Springer, pp. 271–290.

Martínez-Meyer, E. et al. 2013. Ecological niche structure and rangewide abundance patterns of species. – Biol. Lett. 9: 20120637.

Meier E. S. et al. 2010. Biotic and abiotic variables show little redundancy in explaining tree species distributions. – Ecography 33: 1038–1048.

Merow, C. et al. 2013. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. – Ecography 36: 1058–1069.

Merow, C. et al. 2014. What do we gain from simplicity versus complexity in species distribution models? – Ecography 37: 1267–1281.
Moreno-Amat, E. et al. 2015. Impact of model complexity on cross-temporal transferability in Maxent species distribution models: an assessment using paleobotanical data. – Ecol. Model. 312: 308–317.
Muscarella, R. and Uriarte, M. 2016. Do community-weighted mean functional traits reflect optimal strategies? – Proc. R. Soc. B 283: 20152434.
Muscarella, R. et al. 2014. ENMeval: an R package for conducting spatially independent evaluations and estimating optimal model complexity for MAXENT ecological niche models. – Methods Ecol. Evol. 5: 1198–1205.
Papeş, 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. 2007. Species' distribution modeling for conservation educators and practitioners. Synthesis. – American Museum of Natural History, <http://ncep.amnh.org>.
Pearson R. G. and Dawson T. P. 2003. Predicting the impacts of climate change on the distribution of species are bioclimate envelope models useful? – Global Ecol. Biogeogr. 12: 361–371.
Peterson, A. T. 2006. Uses and requirements of ecological niche models and related distributional models. – Biodivers. Inform. 3: 59–72.
Peterson, A. T. et al. 2011. Ecological niches and geographic distributions. – Princeton Univ. Press.
Phillips, S. J. 2008. A brief tutorial on Maxent. – American Museum of Natural History, <http://ncep.amnh.org>.
Phillips, S. J. and Dudík, M. 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. – Ecography 31: 161–175.
Phillips, S. J. et al. 2006. Maximum entropy modeling of species geographic distributions. – Ecol. Model. 190: 231–259.
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.
Prates, I. et al. 2016. Inferring responses to climate dynamics from historical demography in neotropical forest lizards. – Proc. Natl Acad. Sci. USA 113: 7978–7985.
Radosavljevic, A. and Anderson, R. P. 2014. Making better Maxent models of species distributions: complexity, overfitting and evaluation. – J. Biogeogr. 41: 629–643.
Rakotondratsimba, H. and Goodman, S. M. 2014. Technical aspects. – In: Goodman, S. M. and Raherilalao, M. J. (eds), Atlas of selected land vertebrates of Madagascar. Association Vahatra, Antananarivo, Madagascar, pp. 19–34.
Reddy, S. and Dávalos, L. M. 2003. Geographical sampling bias and its implications for conservation priorities in Africa. – J. Biogeogr. 30: 1719–1727.
Royle, J. A. et al. 2012. Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. – Methods Ecol. Evol. 3: 545–554.
Schoener, T. W. 1968. The anolis lizards of Bimini: resource partitioning in a complex fauna. – Ecology 49: 704–726.
Searcy, C. A. et al. 2015. Determinants of size at metamorphosis in an endangered amphibian and their projected effects on population stability. – Oikos 124: 724–731.
Shcheglovitova, M. and Anderson, R. P. 2013. Estimating optimal complexity for ecological niche models: a jackknife approach for species with small sample sizes. – Ecol. Model. 269: 9–17.
Soarimalala, V. and Goodman, S. M. 2011. Les petits mammiferes de Madagascar. – Association Vahatra, Antananarivo, Madagascar.
Soberón, J. M. et al. 2000. The use of specimen-label databases for conservation purposes: an example using Mexican papilionid and pierid butterflies. – Biodivers. Conserv. 9: 1441–1466.
Thuiller, W. et al. 2009. BIOMOD – a platform for ensemble forecasting of species distributions. – Ecography 32: 369–373.
Veloz, S. D. 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. – J. Biogeogr. 36: 2290–2299.
Waltari E. et al. 2007. Locating Pleistocene refugia: comparing phylogeographic and ecological niche model predictions. – PLoS One 2: e563.
Warren, D. L. and Seifert, S. N. 2011. Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. – Ecol. Appl. 21: 335–342.
Warren, D. L. et al. 2008. Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. – Evolution 62: 2868–2883.
Warren, D. L. et al. 2014 Incorporating model complexity and spatial sampling bias into ecological niche models of climate change risks faced by 90 California vertebrate species of concern. – Divers. Distrib. 20: 334–343.
Williams, S. E. et al. 2009. Ecological specialization and population size in a biodiversity hotspot: how rare species avoid extinction. – Proc. Natl Acad. Sci. USA 106: 19737–19741.
Wisz, M. S. et al. 2008. Effects of sample size on the performance of species distribution models. – Divers. Distrib. 14: 763–773.

Supplementary material (Appendix ECOG-02909 at <www.ecography.org/appendix/ecog-02909>). Appendix 1.