"MODEL COMPLEXITY AND VARIABLE SELECTION IN MAXENT NICHE MODELS: ANALYSES FOR RODENTS IN MADAGASCAR"

Peter Galante
CUNY City College

Recommended Citation
Galante, Peter, "MODEL COMPLEXITY AND VARIABLE SELECTION IN MAXENT NICHE MODELS: ANALYSES FOR RODENTS IN MADAGASCAR" (2015). CUNY Academic Works.
http://academicworks.cuny.edu/cc_etds_theses/548

This Thesis is brought to you for free and open access by the City College of New York at CUNY Academic Works. It has been accepted for inclusion in Master's Theses by an authorized administrator of CUNY Academic Works. For more information, please contact AcademicWorks@cuny.edu.
MODEL COMPLEXITY AND VARIABLE SELECTION

IN MAXENT NICHE MODELS:

ANALYSES FOR RODENTS IN MADAGASCAR

Peter J. Galante

Department of Biology, City College of New York, City University of New York, New York, NY 10031, USA

Masters Thesis Committee:

Robert P. Anderson, Ph.D. (advisor)

Ana C. Carnaval, Ph.D.

Michael J. Hickerson, Ph.D.

9 July 2015
Table of Contents

Chapter 1: Comparing two approaches for selecting optimal complexity in ecological niche models: information criteria vs. performance on withheld data.................................3

Chapter 2: Selecting input variables for Maxent models of environmental suitability: what goes in and what gets used?.................................................................21

Acknowledgments..................................................................................................................37

References..............................................................................................................................38

Tables................................................................................................................................44

Figures.................................................................................................................................48

Supplementary Material........................................................................................................51
Chapter 1: Comparing two approaches for selecting optimal complexity in ecological niche models: information criteria vs. performance on withheld data

Abstract

Ecological niche models (ENMs) characterize the relationship between localities where a species is known to occur and the abiotic characteristics of these regions. While widely used, ENMs remain subject to several outstanding issues, including those related to model complexity and violation of modeling assumptions (e.g., representative sampling). Critical in resolving these issues is a better understanding of the effectiveness of model selection techniques. Here, I compare two strategies for optimizing ENMs: an information-criterion approach (AICc) and a sequential approach that assesses model performance on withheld data. I do so for a single species using two datasets, one with all available occurrence records, and the other with spatially filtered occurrence records (expected to reduce the effect of sampling bias). I conduct these experiments making models with Maxent for a species with few occurrence records, the endemic Malagasy rodent *Eliurus majori* (subfamily Nesomyinae), using 19 bioclimatic variables. Candidate models were created across a wide range of complexities. For both datasets, both model-selection techniques chose simpler models than Maxent’s default settings. In the unfiltered dataset, the models selected as optimal by AICc had substantially fewer parameters than those selected by the sequential technique. In contrast, both techniques converged on similar settings when the spatially filtered dataset was used, possibly due to the relative lack of sampling bias present, which better fulfilled important niche modelling assumptions. Nevertheless, the results of each respective selection technique,
and default settings differed between unfiltered and filtered datasets. Qualitative examination of predictions in light of expert knowledge indicated that the selection techniques yielded more realistic models than did the default settings, and those models made with the filtered dataset more closely matched available distributional and natural history information for the species. To reach general conclusions regarding these issues, similar studies should be undertaken with a wide variety of simulated and real species datasets.

**Introduction**

Ecological niche models (ENMs) constitute an important tool in ecology and evolution, but their application is hindered by critical outstanding methodological issues such as selection of optimal levels of model complexity. Correlative ENMs determine the relationship between localities where a species is known to occur, and the abiotic (e.g., climatic) and biotic (e.g. vegetation, species interactions) properties of these locations (Elith et al. 2006). With important assumptions, these models can be projected into other time periods and geographic areas for climate-change and invasive-species studies. Furthermore, they can provide predictions for applications regarding conservation, agriculture, zoonotic diseases, and many other areas of research (Elith & Leathwick 2009).

Although the high performance and ease of use of the ecological niche modeling algorithm Maxent (Phillips et al. 2006) has led to common use, estimating optimal levels of model complexity remains a key outstanding methodological issue. Maxent allows for the use of default settings for factors that affect model complexity (e.g., feature classes and regularization multipliers; Phillips & Dudik 2008). The default settings were selected
as best on average using empirical data spanning a broad range of geographic regions, habitat types, taxonomic groups, number of localities, and sampling biases (Phillips & Dudík 2008). In contrast, users can create models with a wide range of settings for each species, yielding many candidate models to identify the species-specific settings that lead to optimal levels of model complexity. Repeatedly, such species-specific tuning of model settings (also termed “smoothing”) has been shown to result in simpler and substantially better models than those built using default settings (Anderson & Gonzalez 2011; Elith et al. 2010; Muscarella et al. 2014; Radosavljevic & Anderson 2014; Warren et al. 2014; Warren and Seifert 2011).

However, no general consensus yet exists regarding the best way to select optimal complexity—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 evaluating model performance on external (withheld) evaluation data. Regarding the first approach, a substantial number of studies have used information criteria, specifically Akaike’s Information Criterion corrected for small sample size (AICc), in the selection of optimally complex ENMs (Baldwin 2009; Warren & Seifert 2011). Information criteria were not originally developed for machine learning algorithms, and do not fit that paradigm perfectly (Warren & Seifert 2011). Specifically, the degrees of freedom for each model cannot be calculated exactly (Warren et al. 2014). However, AICc can be useful for ENMs because it gives a quantitative measure of model optimization without the use of external evaluation data, balancing model complexity with goodness-of-fit (Guisan & 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, a two recent studies have proposed estimation of optimal complexity based on sequential evaluation criteria, the first of which minimizes overfitting to calibration data and the second maximizing discriminatory ability, sequentially (Radosavljevic & Anderson 2014; Shcheglovitova & Anderson 2013). In this study I implement these particular evaluation metrics, but others are available (Peterson et al. 2011; Warren & Seifert 2011). Although numerous studies implement either AICc or the sequential method to model tuning, to my knowledge, no explicit comparison between the two has yet been conducted.

Additionally, I address one important possible confounding factor: the effects of sampling bias. Although both AICc and the sequential method assume that occurrence data (known species’ localities) derive from unbiased sampling, such an assumption is likely violated in most datasets (Hijmans et al. 2000; Peterson et al. 2011; Phillips et al. 2009). Therefore, to test if either model-selection technique is affected by spatial sampling bias (which likely results in environmental bias; Reddy & Dávalos 2003; Kadmon et al. 2004), I conduct these experiments with two datasets that should reflect different levels of bias (see Boria et al. 2014). Specifically, I 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. To produce the second dataset, I spatially filter the localities, yielding a dataset that should reflect relatively less sampling bias (Veloz 2009; Carroll 2010; Anderson & Raza 2010; Hijmans 2012; Boria et al. 2014).
I construct and compare the models selected by the internal-performance approach (AICc) with those chosen with measures of external performance (sequential criteria). I address these issues with an endemic Malagasy forest-dwelling species known from a fairly small number of unique occurrence records (23), *Eliurus majori* (Soarimalala & Goodman 2011). I do so by creating a suite of models built with a range of settings that affect model complexity (feature classes and regularization multipliers; see Methods). Out of that suite of models, I compare those identified as optimal by the two techniques, as well as with the model produced using Maxent’s default settings. The present experiment addresses three main questions regarding the results of model selection techniques and differently biased datasets:

**Question 1:** How do model complexity and geographic predictions differ between default settings and the two model-selection techniques?

Maxent’s default settings have a tendency to produce overfit models (see above), leading to the following expectations. For each dataset (unfiltered or filtered), I expect that each model-selection technique (AICc or sequential) will identify models that are simpler (fewer parameters) and less overfit (showing lower omission rates) than the one made using default settings. Similarly, I expect the geographic predictions of models selected using the two techniques to differ from 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 sequential)?

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

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

For each selection technique, as well as the default settings, I 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**

I compiled locality information for *Eliurus majori* from museum voucher specimens. IDs were inferred from a phylogeny based on the mitochondrial gene *cytochrome b* in which individuals assigned to *E. majori* were most closely related to those from that clade (with morphological confirmation by specialist S. M. Goodman) than from any other *Eliurus* species (Jansa *et al.* 1999; Jansa, *unpublished data; n = 23* unique localities). To reduce the likely effects of sampling bias in this dataset, I spatially filtered the 23 original localities such that the maximum number of localities was retained. Because of the heterogeneous landscape of this study region and the inferred level of sampling bias across geography (Goodman *et al.* 2014), I used a 10 km filtering distance, resulting in 14 occurrences (see Boria *et al.* 2014). To do so, I used a preliminary version of spThin (Aiello-Lammens *et al.* 2015) in R (R Development Core
Team 2014) 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, I conducted all analyses first with the unfiltered dataset and then with the spatially filtered one.

As environmental predictors, I used 19 bioclimatic variables from WorldClim.org (Hijmans et al. 2005). These data give aspects of temperature and precipitation, and have been shown to produce informative niche models of non-volant mammals (Elith et al. 2006; Jezkova et al. 2009; Anderson & Raza 2010). They are likely relevant for modelling this species, which appears associated with wet, montane conditions (Goodman et al. 2014). 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 & Dudík 2008). I 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 & Raza 2010; Peterson et al. 2011; Anderson 2013, Anderson & Martinex-Meyer 2004). Specifically, I used a bounding box encompassing a 0.5º buffer around the most extreme locality in each of the four cardinal directions. This was done with the 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**

I 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 enforces simplicity by applying penalties for additional parameters included in a model, and higher weights for them (Phillips & Dudik 2008; Phillips *et al.* 2009; Merow *et al.* 2013). Hence, higher regularization protects against overfitting. In particular, I 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, I built models across a range of levels of regularization. By default, Maxent assigns a particular $\beta$ 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 $\beta$ value to increase or decrease the penalties assigned, across all feature classes in concert. Therefore, for each FC combination, I built models across a set of RM values that ranged from 1–5, increasing by increments of 0.25. This resulted in a suite of 68 combinations of FC/RM settings, yielding 68 candidate models.

I made the models in Maxent (version 3.3.3k) using the R package **ENMeval** (version 0.1.1; Muscarella *et al.* 2014, dependencies: rJava, Urbanek 2013; dismo, Hijmans *et al.* 2013; raster, Hijmans 2014). I extracted AICc values, number of parameters (lambda values; see below), a measure of overfitting (omission rate), and
AUC (see below) required to answer my questions. To generate evaluation statistics based on withheld data, I employed the jackknife method of ENMeval (see below). I 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). I disallowed Maxent to sample background pixels in which occurrence localities lay using the “noaddsamplestoBehind” argument of ENMeval. 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 rank.

Additionally, to quantify concordance of resulting predictions in geographic space across the entirety of Madagascar (projecting well beyond the calibration study region; see below) I re-calibrated models in the Maxent graphic user interface. Here, I 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 & Fithian 2013 for assumptions). I then calculated Schoener’s $D$-values (Schoener 1968) for pairs of the resulting predictions in 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, I converted predictions to binary maps according to the 10% training omission-rate threshold of that model using the R package biomod2 (Thuiller et al. 2014). Using these binary maps, I measured the altitudinal range of each prediction, calculated the proportion of pixels
predicted present, and measured binary concordance among predictions using the package `raster` (Hijmans 2014). 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**

I first identified the optimal model using AICc (see Warren & 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 ($\Delta$AICc; the difference between the likelihood of a given model and that of the best model), AICc weights, and evidence ratios. Additionally, to allow comparisons of overfitting with the sequential-optimal models, for model settings selected as optimal by AICc, I obtained the average omission rate using the jackknife procedure and the same threshold as for sequential criteria (see below). AICc (rather than BIC) was used in this experiment because the
available parameter space is massive, and I do not expect that any of my 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).

**Sequential criteria with jackknife**

Next, I identified the optimal model using sequential criteria based on performance on withheld data. To obtain evaluation statistics, I partitioned localities using a jackknife technique, which is most useful for small samples size of localities (small $n$). The jackknife consists of $n$ iterations; in each iteration $n - 1$ localities are used for calibration, and the model is evaluated on the withheld (independent) locality (Pearson 2007; Shcheglovitova & 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, I employed the OR and AUC metrics. ORs (the proportion of evaluation localities omitted) indicate whether a model is overfit to the calibration data. In contrast, AUC gives a relative measure of discriminatory ability across all signal strengths, by quantifying the proportion of instances in which a randomly selected occurrence record ranks higher than a randomly selected background pixel (Peterson et al. 2011). Here, I first identified the models that displayed the lowest average OR. Then, of that subset of models, I chose the one with the highest average AUC score (Shcheglovitova & Anderson 2013). 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. Specifically, I 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; Radosavljevic & Anderson 2014; Shcheglovitova & Anderson 2013). AUC values were similarly averaged across all jackknife iterations, yielding an average AUC for each model setting.

**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 1.1). There was a large difference in the number of parameters used between AICc-optimal and default models in each dataset. The same pattern was apparent for the sequential technique (although weaker for the unfiltered dataset than the filtered one). Likewise, optimal settings for each selection technique consistently led to models with lower overfitting than did 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 & Anderson 2014), those of the default models were far higher, indicating higher levels of overfitting (Table 1.1).

Regarding the geographic predictions, Schoener’s $D$-values as well as binary concordance also matched expectations (Table 1.2). 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 sequential technique and default models show low similarity, although somewhat higher when 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 notably different from the models selected by either technique, and all three models showed small, restrictive areas as suitable. These areas were mostly clustered around known occurrence localities. The model selected by the sequential technique and the model created using default settings both predicted the same elevation range as suitable (454m–2744m), which was more restrictive than the range indicated by the model selected by AICc (171m–2744m). For the filtered dataset, models chosen by the model selection techniques were drastically different from the model created using default settings. The default model was much more restricted to areas immediately around the occurrence localities, yet spanned a larger elevation (171m–2744m), while the other two models showed much larger areal extent, yet were narrower in elevational range (442m–2744m).

**AICc versus sequential technique (Question 2)**

The models selected as optimal using the two model-selection techniques had different levels of complexity, overfitting, and geographic predictions when using the unfiltered dataset, but these measures were similar using the filtered dataset (Table 1.1). The unfiltered dataset led to a large discrepancy in the number of parameters used by
optimal models. In contrast, the models selected using the filtered dataset incorporated very similar numbers of parameters. Within 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 1.2). 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 (see above). 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 (see above).

**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 $D$ ($\text{AICc} = 0.6294$; sequential = 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%; sequential = 93.55%; default = 95.46%; see Table 1.2).

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; sequential: 5.04% unfiltered, 11.42% filtered; default: 4.71% unfiltered, 9.16% filtered).

The models created in this experiment varied substantially in geographic space between spatially unfiltered and filtered datasets (Figure 1.1). For the unfiltered dataset, the three models each predicted a fairly small proportion of the island as highly suitable, with fairly small differences among them. In those models, the particular areas with high suitability fell in eastern portions of the wet 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 predictions from the two selection techniques (AICc and sequential) displayed very high similarity. Each indicated strong prediction throughout wet areas of intermediate and high elevation (including the westernmost known locality for this species and other large extents of the highland plateau 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 wet highland regions, but not to the degree of any of the models made with unfiltered records.
Discussion

Following expectations, the two model-selection techniques produced simpler models (with fewer parameters) than did default settings. Furthermore, the lower omission rates obtained with AICc and the sequential technique indicated lower levels of overfitting than when using default settings. These two model selection techniques also led to more realistic geographic predictions (see below), especially 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 highland plateau 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 wet highland plateau, which includes the westernmost known locality of the species. The elevational ranges indicated by binary optimal models selected using this dataset were varied. The model selected by the sequential technique indicated the same elevation range as the default model, both of
which were more similar to the observed elevation range of this species (875m – 2500m; Soarimalala & 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 the full plateau, including the westernmost locality (see Figure 1.1). Reconstructions of vegetation types in Madagascar indicate that such regions held wet montane forests before extensive anthropogenic deforestation (Rakotondratsimba & 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, I interpret that the 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 behavior of these model-selection techniques depends on the level of sampling bias present, suggesting that AICc may be more robust to departures from the assumption of unbiased sampling.

**Future directions**

These results show trends for a dataset of few occurrence localities, and help set an agenda for future investigations that could lead to more general conclusions and recommendations for the field. Key areas needing research include: larger sample sizes, varying levels of sampling bias, and modifications of these selection techniques. Performing similar experiments on other real species from a variety of taxa and regions, as well as on simulated species with known tolerances and varying levels of niche
complexities—each across a wide range of sample sizes—would help elucidate more general conclusions regarding the performance of these two selection techniques. Complementarily, I also suggest exploring alternative sequential criteria, for example other measures of overfitting—such as various thresholding rules and the use of the difference between calibration and evaluation AUC values (Radosavljevic & Anderson 2014; Warren & Seifert 2011). For example, here the minimum training presence threshold (MTP) was uninformative because all candidate models exhibited the same low average omission rate, yielding no differentiation in omission rates among settings. Such research should shed additional light on the utility and performance of these selection techniques, leading towards general recommendations necessary for the broad implementation of niche models in environmental biology.
Chapter 2: Selecting input variables for Maxent models of environmental suitability: what goes in and what gets used?

Abstract

Ecological niche models (ENMs) attempt to estimate the relationship between occurrence localities of a taxon and the environmental characteristics of those regions. ENMs have many applications, for instance, guiding field expeditions to suitable areas for species known from few localities. However, the utility of such applications is highly dependent upon the accuracy of the model, which in turn depends on key implementation practices. One such issue is the selection of environmental variables. Because these variables may be highly correlated and excessive numbers of input variables may lead to overly complex models (with especially poor transferability across space or time), the issue of reducing dimensionality remains a debated topic in the field. On one hand, regression techniques (e.g., GLM, GAM) assume that predictor variables are uncorrelated, and therefore require researchers to reduce predictor sets to a maximum allowable correlation value. In contrast, machine learning algorithms (e.g., Maxent, Boosted Regression Trees) do not require uncorrelated variables, and rely on penalties against complexity for limiting the inclusion of variables and estimating optimal complexity (and dimensionality) through model-selection procedures (tuning or smoothing). Nevertheless, many recent implementations of machine-learning algorithms reduce dimensionality beforehand and rely on default settings for model-selection penalties to estimate optimal complexity. Here, I address dimensionality for Maxent models, comparing four input predictor sets that vary in the number of variables included.
and the correlations among them. One dataset uses a complete set of 19 bioclimatic and
one geological variable, and a second is reduced, containing only those variables
important in building a previously tuned Maxent model. The third and fourth datasets are
reduced via correlation analysis, where highly correlated variables are removed, at either
a fixed threshold, or further reduced to match the number of variables input in the second
predictor set. I do this for a simulated species in Madagascar, for which I know the
geographic distribution of the niche, allowing me to quantify the accuracy of
reconstructions of the true niche for models made with few localities under each of the
predictor sets. I then apply the best-performing variable-reduction strategy to a lineage of
rare rodents, endemic to Madagascar, to estimate suitable areas and guide future field
surveys. Models tuned with predictor set 2 (the variables taken from a previously tuned
Maxent model) had marginally higher success in reconstructing the virtual niche from
which simulated localities were drawn. These models also had superior performance in
predicting withheld data (omission rate) as well as discrimination (AUC). This approach
created a realistic model indicating many new potential areas for sampling of the real
lineage. Taken together, the results indicate that reducing dimensionality via a tuning
process in Maxent can lead to more accurate model predictions, whereas reducing
predictor sets a priori via variable correlation may remove important and relevant
information about the species’ niche. Furthermore, a two-step tuning approach may prove
useful for appropriately estimating optimal model complexity. Similar experiments
should be conducted for simulated and real species with varied niche complexities and
numbers of available occurrence localities.
**Introduction**

Ecological niche models (ENMs) approximate the relationship between a taxon’s known localities and the environmental conditions of those locations. ENMs have a wide range of applications including conservation assessments and guiding field surveys (Anderson 2013). For such applications, ENMs allow researchers to identify previously undocumented suitable areas in which to sample (Pearson *et al.* 2007; Raxworthy *et al.* 2007). Such insights are especially relevant in the context of modeling species known from few occurrence records (Soberón *et al.* 2000).

Two prevalent families of correlative ENMs include regression-based techniques, and machine learning approaches. Regression techniques such as generalized linear models (GLMs) and generalized additive models (GAMs) assess the relationship between an explanatory variable and its mean response, while assuming that occurrence data come from a given probability distribution (Guisan *et al.* 2002). Regression techniques assume that low autocorrelation exists among input variables (Stewart 1987; see Dormann *et al.* 2013). These methods classically use presence-absence data to fit responses for all variables to be incorporated into the model (Lehmann *et al.* 2002). They were applied early-on for assessing species’ niches and distributions, and have been commonly used by substituting pseudoabsence (Zaniewski *et al.* 2002) data when absences are not available. More recently however, machine learning algorithms such as Maxent and Boosted Regression Trees have been shown to have relatively stronger predictive power (Elith & Leathwick 2009). These latter types of algorithms make no assumptions about the probability distribution of the species’ response to environmental variables. By incorporating presence and pseudoabsence/background data, such models iteratively
make and test predictions on internally withheld data such that not all variables are incorporated into the final model.

Under both families, users are free to incorporate a wide variety of input variables; yet, theory dictates the use of variables that have a driving influence on the species’ niche requirements (Austin 2002; see Anderson 2013 for explanations). Often, these variables are unknown to researchers, who therefore use variables that likely act as proxy measures and can be useful through correlations with the driving variables. Ultimately, many variables are input, some of which may be highly correlated. For both approaches, overfitting to calibration data (resulting in poor generality) can be lessened by reducing dimensionality of (number of) input variables (Rushton et al. 2004; Elith et al. 2011; Merow et al. 2013). Additionally, collinearity (high correlation) of input variables can cause computational problems for some techniques, as well as complications in interpreting model output for those studies desiring hypotheses regarding explanation (which variables may cause the species’ response; Dormann et al. 2013).

Regression modeling and machine learning typically approach the issue of variable correlations and dimensionality differently. In regression-based modeling, collinearity among input variables is problematic as it violates modeling assumptions (Guisan & Thuiller 2005). Models are created under the assumption that predictor sets are composed of variables with low levels of correlation often assessed through pairwise Pearson’s product-moment correlation coefficients (Guisan & Zimmermann 2000). In contrast, machine learning algorithms (e.g., Maxent; Phillips et al. 2006) can receive correlated variables, but are still subject to pitfalls regarding over-complexity. For
Maxent, over-complexity hinders model performance when complex feature classes are applied with large sets of input variables (Merow et al. 2013). However, these methods select variables based on the explanatory information gained by including them, balanced against a cost penalty (termed $L_1$ regularization) applied for increasing complexity (Guisan & Thuiller 2005). This $L_1$ regularization reduces variable coefficients (for some variables to zero), causing some variables to not be included in the final model (Elith et al. 2011). An equivalent penalty for linear features (typically called a lasso) also has been applied to regression-based techniques, but is not generally implemented for such studies; rather, multimodel inference using information criteria is sometimes used for variable selection, after reducing variables before modeling (Burnham & Anderson 2004; Symonds & Moussalli 2010). Nonetheless, despite the lack of formal assessment of the effect of variable correlation within a machine-learning environment, many studies using Maxent follow the regression-based approach and measure variable correlation to remove highly correlated variables using arbitrary thresholds before inputting variables into the algorithm (Kramer-Schadt et al. 2013; Lozier et al., 2009; Syfert et al. 2013; see Merow et al. 2013). Typically, researchers use proxy variables because the true driving variables are unknown or unobtainable, but this may not be appropriate for two reasons. First, for both regression-based techniques and machine learning algorithms, reducing input variables via correlation analysis potentially can remove informative and biologically relevant data. Second, this step may not be necessary—or appropriate—for machine-learning algorithms given the internal variable-selection approach implemented in them.

To address these issues, I compare various strategies for assessing variable relationships: two that take variable correlations into account before inputting predictors
into the algorithm, and two that do not. I assess the performance of each strategy for a simulated species, and then implement the best-performing one on a real lineage, of which we known little. Therefore, I conduct experiments using a simulated species with a relatively small sample size ($n = 15$), then apply the highest-performing strategy to a real lineage of morphologically similar sister species known from very few unique occurrence localities (total $n = 7$).

**Methods**

**Simulated species**

I created a simulated niche using the virtualspecies (Leroy et al. 2015) package in R (R Development Core Team 2014). I created this niche such that it had a geographic distribution somewhat similar to that of the real lineage in Madagascar (see below). With the assumption that they will be highly correlated with many other bioclimatic variables, I designated virtual response curves for two variables available from worldclim.org (Hijmans et al. 2005). Specifically, I assigned normal distribution response curves to Annual Mean Temperature and Annual Precipitation (mean of 26.0°C, standard deviation of 50°C; mean of 1500 mm, standard deviation of 500 mm, respectively). In addition, I created a categorical response for a geological variable (Du Puy & Moat 1996) of parent bedrock of Madagascar, apparently an important driver of the distribution of the real lineage (Soarimalala & Goodman 2011). Limestone was given a suitability of 0.690, sandstone a suitability of 0.310, and all other soil types were assigned a suitability value of 0.000. I then translated this environmental niche into a simulated geographic distribution and created a map indicating suitability for the
simulated species across Madagascar. It is important to note that because the virtual niche does not include biotic interactions, and I did not include the effects of any dispersal barriers (or minimal areal requirements of any patch), the species’ distribution can be interpreted as in equilibrium with the climatic determinants.

Localities were drawn from this virtual distribution so that they were proportional to the squared suitability values within the raster. Squaring the pixel values added discrimination between highly suitable pixels and low suitability pixels. In this way, I drew 100 unique datasets of 15 localities each for use in this experiment. Due to the randomized nature of the sampling of localities, I assume that these datasets have no sampling bias, but they differ from each other due to randomness, which can lead to differences in niche models (especially for small sample sizes; Boria et al. 2014)

**Real mitochondrial lineage**

This lineage is comprised of two closely related (sister) species of the genus *Eliurus*: *E. antsingy* and *E. carletoni* (hereafter; *ant-carl*). I used georeferenced specimen records from a phylogeny inferred through cytochrome *b* sequencing in which genetic divergence between these two named forms is less than that found within of other single species in the genus *Eliurus* (Jansa, unpublished data). These two species are both known to have strong associations with low-elevation dry forests, and with sedimentary bedrock (preference for limestone, but also associated with sandy soils; (Rakotoarisoa et al. 2010; Goodman et al. 2014). Each is known from very few localities (3 and 5, respectively). Often, field sampling techniques are biased in geographic space, which can lead to biases in environmental space (Kadmon et al. 2004). Therefore, to reduce likely artefactual
spatial autocorrelation among localities, I spatially filtered all localities using the \textit{spThin} (version 0.1.0; Aiello-Lammens \textit{et al.} 2015) package in R such that no two localities were closer than 10 km (see Boria \textit{et al.} 2014), leading to a final dataset of 5 localities.

\textit{Predictor variable sets}

Sets of predictor variables were determined for the virtual niche in four ways. Predictor variables used in this experiment were the 19 continuous bioclimatic variables and the categorical variable of geology. For the simulation, I used the entire island of Madagascar as the study region. I then determined the pairwise correlations among all predictor variables using Pearson’s product-moment correlation coefficients. Then, I ranked them according to three sequential criteria (as needed to break ties). First, I ranked variables by highest pairwise correlation among the set. Subsequently, I determined the highest average correlation of every variable with all other variables. Lastly, I determined the average correlation remaining in the entire set of variables after the exclusion of each variable as it is removed from the set. Variables were then removed according to the resultant ranking, from highest to lowest correlations.

Predictor variables were subset four different ways:

\textit{All variables}: Predictor variables used in this set include the 19 continuous bioclimatic variables as well as the categorical layer of geology.

\textit{Important variables}: This second set of predictor variables included only those variables that show a contribution to training gain in the optimally tuned model from \textit{All variables} (see \textit{Model Tuning}, below). This method is similar to Wollan \textit{et al.} 2008 and Synes and Osborne 2011; however,
because of a second round of tuning, not all of the variables in this set were incorporated into the final model.

**Least-correlated variables**: This predictor set is a subset of all predictor variables, where variables were removed until the overall correlation coefficient of the variable matrix was below 0.65. The 0.65 threshold is a conservative value found in the literature (Lozier et al. 2009; Syfert et al. 2013; Kramer-Schadt et al. 2013).

**Least-correlated control**: This predictor set used the same number of variables that were included in the **Important variables** set; however, they were selected from the same ranked list of least correlated variables used by **Least-correlated variables**. For each simulated dataset, as the number of variables included of **Important variables** changed, so did the correlation in this predictor set. This set is designed to allow for appropriate comparisons with **Important variables**.

**Model tuning**

Models were tuned for each dataset from each of the variable sets using a modified function from the R package **ENMeval** (version 0.1.1; Muscarella et al. 2014). The modification to the **evaluate** function changed the output results table such that it listed summary statistics and percent contribution of predictor variables (either an increase or decrease in regularized gain during model-building iterations). I tuned the models across a wide range of complexity by varying two Maxent settings: feature classes (FCs), and regularization multiplier (RM; L₁ regularization, see above). FCs
control the shape of the modeled response for each input variable, while increasing RMs apply stronger penalties to parameters included in the model thus enforcing simplicity (Phillips & Dudík 2008; Phillips et al. 2009; Merow et al. 2013). FC combinations used in this experiments span a wide range of complexities reasonable for exploration for the small sample sizes employed in this study: Linear (L); Linear and Quadratic (LQ); Hinge (H); and Linear, Quadratic, and Hinge (LQH; Phillips & Dudík 2008; Shcheglovitova & Anderson 2013; Radosavljevic & Anderson 2014). For each FC combination, I used a range of RM values from 0.5 to 4.0 in stepwise increments of 0.5. This resulted in a total of 32 candidate models for each predictor set of input variables, for each simulated dataset. Because ENMeval can only tune models with at least two predictor variables, those simulated datasets that yielded optimal models with only one variable were removed from analyses. Additionally, to standardize model comparisons, I ensured that all models were tuned using a standardized set of background points (n = 10,000) drawn randomly from the study region.

Due to the small sample sizes employed here, I chose the jackknife method of model evaluation (Pearson et al. 2007; Shcheglovitova & Anderson 2013). To avoid models with obvious overfitting, candidate models with more parameters than localities were removed from all analyses. Similarly, models with no parameters were also removed. Optimal models were then selected using sequential criteria obtained through averaging test statistics for withheld localities across jackknife replicates (Shcheglovitova & Anderson 2013). The first criterion (omission rates) aimed to minimize overfitting, while the second (AUC; see below) maximized the ability of the model to discriminate between a presence locality, and a randomly selected background pixel. Omission rates,
calculated using the 10% calibration omission threshold (OR), reflect overfitting to calibration data (when over 10% test omission), while the Area Under the Curve of the Receiver Operator Characteristic plot (evaluation AUC; hereafter, AUC) is a threshold-independent measure of the discriminatory ability of the model. Candidate models that exhibited no discrimination (AUC value of 0.5 or less) or had an omission rate of zero (did not omit any localities) were removed from further analyses. The latter models (actually achieving omission rates better than expected theoretically) usually corresponded to non-informative models with little no parameters, which predict all of the study region as equally present. I selected candidate models exhibiting the lowest OR, then, as the secondary criterion of that subset of candidates, I selected, as optimal, the model with the highest AUC.

Model comparisons

By using simulations, I was also able to measure Maxent’s ability to reconstruct the virtual niche for each predictor set. To do this, I compared the niches is geographic space. Specifically, I compared the logistic output for the optimal models from each predictor set to the raster of the simulated niche (for the squared pixel values) using an approach that compares continuous raster surfaces: Schoener’s D-value (Schoener 1968; see Warren & Seifert 2011). To perform this analysis, all final optimal models were created using the best-performing settings and all spatially filtered localities, and then projected to the study region using the logistic output of Maxent. I then compared that with the virtual niche (by calculating D) using the R package dismo (Hijmans et al. 2013). To examine if D-values from each predictor set were significantly different, I
made six pairwise comparisons of $D$-values between predictor sets using Wilcoxon’s Signed-Rank Tests, with a family-wide Bonferroni correction for multiple comparisons. These $D$-values were further compared visually using a boxplot in R (R Development Core Team 2014). I then calculated a 5–95% range of $D$-values to gauge the range of each dataset while removing extreme values. Furthermore, I calculated evaluation-OR and AUC values (averaged over the 15 jackknife iterations). Then, I calculated the proportion of times that each predictor set better reconstructed the virtual niche than All variables (higher Schoener’s $D$-value), achieved lower measures of overfitting (lower OR), and showed better discriminatory ability (higher AUC).

**Real mitochondrial lineage**

Similar methods were used for the real lineage, with one distinction. I made a custom study region for *ant-carl* designed for closer approximation of an important modeling assumption regarding dispersal barriers (Anderson 2013; Anderson & Raza 2010; Peterson *et al.* 2011). To do so, I created a study region of a bounding box encompassing a 0.5º buffer around the most extreme locality in each of the four cardinal directions. The predictor set approach with the highest performance for the simulated species was used for the real lineage.

**Results**

**Simulated species**

Comparison of optimal models from each predictor set with the virtual niche using Schoener’s $D$-value revealed several clear trends. Optimal models created using
predictor sets that did not perform an *a priori* assessment of variable correlation (*All variables*, and *Important variables*) performed better than those that did. Optimal models created using *All variables* and *Important variables* had highest $D$-values (when comparing with the “true” virtual niche; Figure 2.2, Table 2.1). Models from *Important variables* were better at reconstructing the virtual niche slightly more often than those from *All variables*; however, $D$-values were not significantly higher (Table 2.2). $D$-values from both datasets that assessed correlation (*Least-correlated variables*, and *Least-correlated control*) were lower than those from *All variables*, yet only those from *Least-correlated control* were significantly lower than models from any other predictor set. The median $D$-values from *All variables* and *Important variables* were remarkably similar, yet the values from *Important variables* showed a much smaller 5-95% range. The 5-95% range of $D$-values from *Least-correlated* was remarkably similar to *All variables*, while the *Least-correlated control* range was much larger.

These trends were largely echoed by the statistics that evaluated performance on withheld data. Optimal models from the *Important variables* predictor set either performed better than or equal to those of the *All variables* predictor set, while *Least-correlated variables*, and *Least-correlated control* were more commonly performed worse (Table 2.1). Although omission rates from the *Important variables* set were only rarely better than those of *All variables*, they were never worse. In contrast, the *Least-correlated variables* and *Least-correlated control* predictor sets produced models that had lower omission rates more often than *All variables*; importantly, they were more prone to error, including several instances of being worse (4 times and 6 times, respectively; Table 2.1). *Important variables* exhibited better AUC values than *All
variables an overwhelming majority of times (84%), but the two predictor sets that were reduced by correlation analysis showed poorer AUCs more often, when compared with optimal models from All variables (Table 2.1).

Real lineage

Based on the results for the simulated species, model complexity was optimized (tuned) for the real lineage using the Important variables approach. For this lineage, that led to input of geology, temperature annual range, and precipitation of the driest month. The optimal model from this predictor set identified only geology and temperature annual range as important for model creation. Two localities were omitted at the 10% OR logistic threshold, resulting in a high OR of 0.4. The evaluation-AUC was 0.936. Geographic areas of high suitability follow the limestone geology of north-western Madagascar, with the areas of very high suitability lying at the northern extreme of these geological formations (Figure 2.2).

Discussion

The comparison of the four variable-selection approaches implemented for the simulated species indicated that using Maxent’s regularization and tuning to identify optimal complexity showed the highest performance (with assumptions when projecting to different regions or time periods; Table 2.1). Furthermore, between the two former techniques, using Maxent to pre-select predictor variables (Important variables) resulted in slightly better reconstructions of the virtual niche (higher $D$-values) than All variables, as well as moderately better evaluation statistics (Figure 2.2, Table 2.1). On the other
hand, the *Least-correlated control* dataset led to significantly worse reconstructions of
the virtual niche than any other predictor set. Reducing variables via correlation analysis,
as advised for regression-based modeling, clearly did not perform well here. It may
remove relevant and important information about the species’ niche. In contrast, the
trends found in the present experiments indicate that reducing the input dataset to those
variables that are important to modeling the species (via a preliminary round of
modeling) may more closely match modeling assumptions of using only relevant input
variables, as well as to minimize complexity and protect generality in predictions (Austin,
2002; see Anderson, 2013 for importance of generality).

Despite only having 5 localities, the optimally tuned model for the real lineage
created from *Important variables* shows a fairly reasonable geographic prediction that
follows the suspected necessary environmental requirement of Karst and sandstone
gеology (Soarimalala & Goodman 2011). Although Goodman *et al.* (2014) suggested that
dry to humid-dry forests may be important to the biology of the species, evidently the
data conferred by temperature annual range and geology were sufficiently correlated with
other variables to maintain niche signal. The geographic region predicted as suitable for
*ant-carl* indicates a much larger suitable area than is currently known for either species
(Goodman *et al.* 2014), including several disjunct regions of suitable conditions. This
model also indicates several fairly large areas of currently forested regions (not shown:
Rakotondratsimba & Goodman 2014) in which this species has not been currently
sampled, suggesting that these areas should be prioritized in future sampling efforts for
this rare lineage of Malagasy endemic rodents.
Here, the reduction of input variables via correlation analysis did not improve model performance in a virtual species known from few localities. In contrast, using Maxent to identify variables that are important to modeling the species’ niche may correctly identify the niche (or at least variables and conditions associated with it) more often than using all variables. This reduction in dimensionality of input variables through a two-step tuning approach may adequately control the overly-complex models that Maxent can create, while still maintaining important variables (Merow et al. 2013; Yu & Liu 2003).

Future studies should examine the generality of these results. For example, similar experiments could be conducted for regions (and input variable types) where the predictor variables have varying levels of correlation. Furthermore, and likely of great importance, these analyses should be conducted for a variety of sample size of occurrence localities. Performing these analyses on varying geographic regions, variable-correlation structures, and sample sizes, may provide more realistic expectations of model performance.
Acknowledgments:

I thank the members of Ecology, Evolution and Behavior labs at The City College of New York for providing a great environment for learning and support. Jason L. Brown put up with my many questions, and provided help in this research. Robert A. Boria, Maria Gavrutenko, Beth Gerstner, Steven M. Goodman, Sharon A. Jansa, Jamie M. Kass, and Mariano Soley-Guardia provided diligent feedback, patience, and insight on one or more ideas presented in this thesis. I further thank Sharon and Steve for their collaborative efforts and providing locality information used in this research. Robert Muscarella and Diego Alvarado-Serrano have been invaluable as teachers as well as collaborators. Lastly, I would like to thank Robert P. Anderson for his consummate professionalism and patience, and making every moment a chance for me to learn. This research was made possible by funding from the National Science Foundation (DEB-1119915 to RPA).
References:


Tables:

**Table 1.1**: Summary of optimal tuning experiments for Maxent models of the Malagasy rodent *Eliurus majori*. Results are provided for two model-selection techniques (AICc and sequential criteria) as well as the default settings, for two datasets (unfiltered and filtered localities). Settings, number of parameters (non-zero lambda values), and the omission rate (OR) for evaluation localities is provided for all experiments.

<table>
<thead>
<tr>
<th></th>
<th>Unfiltered (n = 24)</th>
<th>Filtered (n = 13)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Settings</td>
<td>λ</td>
</tr>
<tr>
<td><strong>AICc</strong></td>
<td>LQ1.5</td>
<td>6</td>
</tr>
<tr>
<td><strong>Sequential</strong></td>
<td>H1.5</td>
<td>14</td>
</tr>
<tr>
<td><strong>Default</strong></td>
<td>LQH1</td>
<td>17</td>
</tr>
</tbody>
</table>
Table 1.2: Summary of comparisons of continuous and binary model predictions (logistic output) 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 sequential criteria) as well as the default settings, for each of two datasets (spatially unfiltered and filtered localities).

<table>
<thead>
<tr>
<th>Continuous Comparisons</th>
<th>Sequential</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) UNFILTERED DATASET</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AICc</td>
<td>0.8172</td>
<td>0.7915</td>
</tr>
<tr>
<td>Sequential</td>
<td>---</td>
<td>0.8522</td>
</tr>
<tr>
<td>(B) FILTERED DATASET</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AICc</td>
<td>0.9407</td>
<td>0.7494</td>
</tr>
<tr>
<td>Sequential</td>
<td>---</td>
<td>0.7103</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Binary Comparisons</th>
<th>Sequential</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>(C) UNFILTERED DATASET</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AICc</td>
<td>96.24%</td>
<td>96.07%</td>
</tr>
<tr>
<td>Sequential</td>
<td>---</td>
<td>98.65%</td>
</tr>
<tr>
<td>(D) FILTERED DATASET</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AICc</td>
<td>99.49%</td>
<td>95.65%</td>
</tr>
<tr>
<td>Sequential</td>
<td>---</td>
<td>95.24%</td>
</tr>
</tbody>
</table>
Table 2.1: Results from simulated datasets drawn from the virtual niche created on Madagascar, showing differences among four variable-selection approaches. Proportion of instances each approach performed better than *All variables* is provided. Additionally, the median value and range where 90% of the data lies are provided for three evaluation statistics: Schoener’s $D$-value, evaluation omission rate, and evaluation-AUC. For omission rate (OR), the proportion of instances that ORs were equal to, and worse than *All variables* (middle/end, respectively) also appears.

<table>
<thead>
<tr>
<th>Proportion of instances that $D$-values performed better than <em>All variables</em></th>
<th>All variables</th>
<th>Important variables</th>
<th>Least-correlated variables</th>
<th>Least-correlated control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median $D$-value</td>
<td>0.5909</td>
<td>0.5905</td>
<td>0.549</td>
<td>0.4718</td>
</tr>
<tr>
<td>5% to 95% range</td>
<td>0.376–0.756</td>
<td>0.409–0.640</td>
<td>0.315–0.751</td>
<td>0.240–0.751</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Proportion of instances that ORs were better than <em>All variables</em></th>
<th>---</th>
<th>7%</th>
<th>34%</th>
<th>27%</th>
</tr>
</thead>
<tbody>
<tr>
<td>worse than <em>All variables</em></td>
<td>---</td>
<td>0%</td>
<td>4%</td>
<td>5%</td>
</tr>
<tr>
<td>Median OR</td>
<td>0.1333</td>
<td>0.1333</td>
<td>0.1333</td>
<td>0.1333</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Proportion of instances that AUC values were better than <em>All variables</em></th>
<th>---</th>
<th>84%</th>
<th>49%</th>
<th>41%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median AUC</td>
<td>0.9409</td>
<td>0.9481</td>
<td>0.941</td>
<td>0.9314</td>
</tr>
<tr>
<td>5% to 95% range</td>
<td>0.902–0.968</td>
<td>0.909–0.971</td>
<td>0.834–0.967</td>
<td>0.749–0.964</td>
</tr>
</tbody>
</table>
Table 2.2: Summary of comparisons of a Malagasy simulated species using four sets of input variables. Results are provided for p-values, with significance (asterisks) assessed through a Bonferroni correction, resulting from Wilcoxon’s Signed-Rank Test of pairwise comparisons of resultant Schoener’s $D$-values from four predictor sets using 100 simulated datasets drawn from virtual niche on Madagascar.

<table>
<thead>
<tr>
<th></th>
<th>Important variables</th>
<th>Least correlated control</th>
<th>Least correlated</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All variables</strong></td>
<td>0.9165</td>
<td>0.033</td>
<td>8.37x$10^{-9}$*</td>
</tr>
<tr>
<td><strong>Important variables</strong></td>
<td>---</td>
<td>0.021</td>
<td>2.53x$10^{-6}$*</td>
</tr>
<tr>
<td><strong>Least correlated</strong></td>
<td>---</td>
<td>---</td>
<td>0.001*</td>
</tr>
</tbody>
</table>
Figures:

Figure 1.1: 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 data (middle), and default settings (right). Plotted localities (black dots) represent unfiltered or filtered occurrence records for the corresponding row.
Figure 2.1: Logistic output of the Maxent model of the lineage *ant-carl* (comprised of *Eliurus antsingy* and *E. carletoni*), projected to the study region used for modeling. The model settings were tuned using sequential criteria and the *Important variables* predictor variable set. Warmer colors indicate those areas of higher suitability for the lineage.
Figure 2.2: Boxplot of Schoener’s $D$-values of optimal models selected for all 100 simulated datasets of a simulated species on Madagascar using four different predictor sets. Letters indicate significant differences as determined through a Wilcoxon’s test with a Bonferroni correction. Notches illustrate approximately 95% confidence intervals around the median. Boxes represent first and third quartile of $D$-values. Whiskers represent the most extreme data point that is no more than 1.5 times the length of the box. Dots represent statistical outliers.
Supplementary Material:

**Supplementary Table 2.1:** Bioclimatic and Geology input variables ranked by descending correlation according to three sequential criteria for both the virtual niche (all of Madagascar) and the real lineage (*ant-carl*; buffered bounding box) for experiments comparing models made by four predictor sets. Asterisks indicate those variables included in the *Least-correlated variables* predictor set.

<table>
<thead>
<tr>
<th>Simulated Species</th>
<th>Real Lineage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean temperature of warmest quarter</td>
<td>Mean diurnal range</td>
</tr>
<tr>
<td>Precipitation of driest month</td>
<td>Isothermality</td>
</tr>
<tr>
<td>Precipitation of driest quarter</td>
<td>Min temperature of coldest month</td>
</tr>
<tr>
<td>Mean temperature of coldest quarter</td>
<td>Mean temperature of driest quarter</td>
</tr>
<tr>
<td>Precipitation of wettest quarter</td>
<td>Annual mean temperature</td>
</tr>
<tr>
<td>Annual mean temperature</td>
<td>Precipitation of coldest quarter</td>
</tr>
<tr>
<td>Mean diurnal range</td>
<td>Annual precipitation</td>
</tr>
<tr>
<td>Max temperature of warmest month</td>
<td>Precipitation of wettest quarter</td>
</tr>
<tr>
<td>Min temperature of coldest month</td>
<td>Precipitation seasonality</td>
</tr>
<tr>
<td>Mean temperature of driest quarter</td>
<td>Max temperature of warmest month</td>
</tr>
<tr>
<td>Precipitation of coldest quarter</td>
<td>Temperature annual range</td>
</tr>
<tr>
<td>Annual precipitation</td>
<td>Mean temperature of coldest quarter</td>
</tr>
<tr>
<td>Temperature seasonality*</td>
<td>Precipitation of driest month*</td>
</tr>
<tr>
<td>Precipitation of warmest quarter*</td>
<td>Temperature seasonality*</td>
</tr>
<tr>
<td>Temperature annual range*</td>
<td>Precipitation of warmest quarter*</td>
</tr>
<tr>
<td>Precipitation seasonality*</td>
<td>Mean temperature of warmest quarter*</td>
</tr>
<tr>
<td>Geology*</td>
<td>Mean temperature of wettest quarter*</td>
</tr>
<tr>
<td>Isothermality*</td>
<td>Precipitation of driest quarter*</td>
</tr>
<tr>
<td>Mean temperature of wettest quarter*</td>
<td>Precipitation of wettest month*</td>
</tr>
<tr>
<td>Precipitation of wettest month*</td>
<td>Geology*</td>
</tr>
</tbody>
</table>