2011

Using geographically structured evaluations to assess performance and transferability of ecological niche models for species with many occurrence records; a test using the Caribbean spiny pocket mouse, Heteromys anomauls

Aleksandar Radosavljevic
CUNY City College

How does access to this work benefit you? Let us know!
Follow this and additional works at: http://academicworks.cuny.edu/cc_etds_theses

Part of the Biology Commons

Recommended Citation
Radosavljevic, Aleksandar, "Using geographically structured evaluations to assess performance and transferability of ecological niche models for species with many occurrence records; a test using the Caribbean spiny pocket mouse, Heteromys anomauls" (2011).
CUNY Academic Works.
http://academicworks.cuny.edu/cc_etds_theses/59

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.
Using geographically structured evaluations to assess performance and transferability of ecological niche models for species with many occurrence records: a test using the Caribbean spiny pocket mouse, *Heteromys anomalus*

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

By: Aleksandar Radosavljevic

Thesis Advisor: Robert P. Anderson, Ph.D.

30 December 2010
Introduction

Challenges in evaluating ecological niche models

In recent years, many techniques for modeling species’ niches and distributions have been developed and applied widely (Guisan and Zimmermann, 2000; Guisan and Thuiller, 2005). Whereas some studies aim to predict the species’ potential distribution, others desire models of the species’ current true distribution. Although the terms have not been used consistently in the literature, the former kind of research often has been called “ecological niche modeling” (or some variation thereof), and the latter “species distribution modeling”. While we focus primarily on the niche-modeling paradigm, many of same principles also are relevant to species distribution modeling.

Ecological niche modeling uses occurrence records of a species together with climatic or other environmental variables to produce a model of the species’ niche in examined dimensions of environmental space. The model is then applied to geographic space to approximate the species’ potential distribution (Peterson, 2003). In contrast to techniques that use data regarding both presences and absences of the species, which are available only for relatively few species (Soberón and Peterson, 2004), some methods require only presence records (often from natural history museums and herbaria; Elith et al., 2006; Newbold, 2010). Most such methods compare environmental conditions from localities of known presence to a sample of the environments available in the study region (generally via a background or pseudoabsence sample; Graham et al., 2004; Phillips et al., 2009). Because of the wide availability of electronic
climatic and occurrence data, presence-only models have become pervasive in environmental biology (Kozak et al., 2008).

Despite notable exceptions, application of such models has outpaced conceptual and methodological research regarding model evaluation (Araújo et al., 2005b; Lobo et al., 2008; Veloz, 2009). Of particular concern is the acquisition and use of truly independent evaluation data. Most studies have evaluated model performance based on random partitioning of occurrence data into calibration and evaluation datasets (split-sample approach of Guisan and Zimmermann, 2000; e.g., Anderson et al., 2002a; Hernandez et al., 2006; Raxworthy et al., 2007; Jezkova et al., 2009), but this tactic has proven problematic (Araújo et al., 2005b). First of all, because calibration and evaluation localities may lie close to each other, evaluation localities are not truly independent of those used to calibrate the model and therefore do not provide realistic tests of model performance (due to spatial autocorrelation of the environment; Veloz, 2009). Secondly, any environmental bias present in the original dataset (which can result from geographic biases associated with frequent sampling near roads, rivers, and population centers; Reddy and Dávalos, 2003; Hortal et al., 2008; Loiselle et al., 2008; Boakes et al., 2010) will be preserved in both calibration and evaluation datasets and can affect model calibration adversely (Wintle et al., 2005; Araújo and Guisan, 2006). Hence, random partitioning cannot detect any overfitting to environmental biases. Overfitting occurs when a model fits the calibration data too closely (in environmental space) and, therefore, fails to predict independent evaluation data
accurately (see Materials and methods). Because of the lack of independence between calibration and evaluation datasets random partitioning only can detect overfitting to random noise present in the calibration dataset. This shortcoming also leads to inflated estimates of performance (Veloz, 2009).

Overfitting to noise or to bias compromises the generality of the model (its ability to predict independent data), reducing its validity and its utility in many applications. Lack of generality is especially problematic for studies that require “transferability” when applying a model to another region or time period (Randin et al., 2006; Peterson et al., 2007; Phillips, 2008). Such applications include the study of invasive species (Peterson, 2003; DeVaneay et al., 2009) and the effects of climatic changes on species distributions (Araújo et al., 2005a; Hijmans and Graham, 2006; Fitzpatrick et al., 2008), as well as the development of management plans based on model results (e.g., Bradley et al., 2010). Because investigators typically lack past occurrence data and never have future occurrence data, evaluation across space has been proposed as the most reasonable approach for achieving realistic evaluations of model transferability (Araújo and Rahbek, 2006).

Transferring a model to another region or time period requires that the species’ response be stationary (Osborne and Suárez-Seoane, 2002). A process is considered stationary if the statistics that define it and that are measured within any subset accurately describe the entire dataset (Osborne et al., 2007). In the context of niche modeling, stationarity requires the following assumptions. First, populations across the range of the species should not differ
in inherited niche characteristics (Murphy and Lovett-Doust, 2007); similarly, cross-time projections assume no niche evolution (Pearson and Dawson, 2003; Nogués-Bravo, 2009). Additionally, the biotic interactions should not differ between the two regions or time periods; often, this will be difficult or impossible to assess; see Anderson et al., 2002b. Furthermore, the second region (or time period) should not include abiotic environments not available in the calibration region. If the latter assumption is violated, causing truncated response curves in the calibration region (Thuiller et al., 2004; Williams and Jackson, 2007), additional assumption/s are required in order to make a prediction in such cells of the projection region (Anderson and Raza, 2010). With these caveats, a researcher can proceed to evaluate model generality and transferability by cross-space evaluations.

In addition to the selection of appropriate evaluation data, another outstanding yet critical issue in ecological niche modeling is the use of default model settings rather than those that provide optimal performance for the species, occurrence localities, study region, and environmental data at hand (Phillips and Dudík, 2008). Building optimal models requires achieving an appropriate balance between simplicity and complexity (avoiding overfitting). Tuning, or “smoothing,” involves varying model parameters to achieve an optimal level of performance; these optimal settings may or may not differ from the default ones. This approach shows promise for improving model performance, particularly if transferable models are desired (Elith et al., 2010; Anderson and Gonzalez, unpublished data). Ideally, specialists should examine model
predictions in geography to confirm that the settings selected as optimal based on quantitative evaluations indeed produce the geographic predictions that correspond most closely to reasonable distributional patterns (based on expert knowledge of the vegetation and habitat types that the species inhabits).

**A proposed evaluation paradigm**

In this study, we propose a novel variation of $k$-fold cross-validation in the context of tuning experiments and assessments of model transferability. In $k$-fold cross-validation (sometimes termed $k$-fold cross partitioning), occurrence localities are divided randomly into $k$ bins (subsets), each of equal sample size (Boyce et al., 2002; Lehmann et al., 2002). Models are then built in an iterative manner, using $(k - 1)$ bins for calibration, with the remaining bin withheld for model evaluation. This is repeated until all bins have been used once for evaluation (i.e., until $k$ models are produced; essentially an $n - 1$ jackknife of bins). The evaluation criterion used in the study (see Materials and methods) can then be averaged over all of the iterations. This method has the advantage (over the standard random split-sample approach) that every occurrence record is used for calibration in $(k - 1)$ models and for evaluation in one model. It also allows for examination of model variability as a consequence of different subsets of the occurrence data being used for model calibration. However, this strategy retains the general drawback of all random split-sample approaches: any geographic biases present in the overall dataset are retained in both calibration
and evaluation datasets—precluding the detection of overfitting to bias and leading to overestimates of performance (see above).

Therefore, following the call by Araújo and Rahbek (2006) for cross-space evaluations of transferability, we propose a modification of \( k \)-fold cross-validation where localities are spatially (geographically) segregated into bins. Each bin, in turn, provides independent evaluation data, allowing evaluations capable of detecting overfitting to bias (in addition to overfitting to noise). This should represent a more rigorous and more realistic model evaluation than traditional \( k \)-fold cross-validation with random bins. Furthermore, models that are capable of predicting evaluation localities accurately based on this strategy should be more general and, hence, more successful under temporal or spatial transferal. However, if any uniform environmental bias in sampling corresponding to the geographic bias in sampling cuts across the geographic bins, this approach will not be able to detect any overfitting to it.

Unfortunately, this approach artificially violates one aspect of the assumption of stationarity of the species’ response (see above), requiring another modification. Since many modeling algorithms use a background or pseudoabsence sample from the study region in model calibration, that region should not include areas where the species is absent because of dispersal limitations or biotic interactions (Anderson and Raza, 2010). This is because background pixels drawn from suitable environments in such regions provide a false negative signal that interferes with successful modeling of the species’ requirements. Similarly, selecting calibration localities from only some portions of
the study region (essentially inserting an artificial bias into the system, as in the geographically structured $k$-fold cross-validation proposed here) mimics the natural processes of dispersal limitation and geographic heterogeneity in biotic interactions that can cause a species to inhabit less than its potential distribution (Anderson and Raza, 2010: 1389). Hence, theory suggests that when employing geographically structured $k$-fold cross-validation (or any other geographically structured data-partitioning scheme; e.g., Peterson et al., 2007) background or pseudoabsence data should not be drawn from areas from which known localities were excluded in model calibration.

Therefore, in the present experiments, we employ three approaches, each a variation on $k$-fold cross-validation. To allow comparison with the conventional implementation of the technique, we first conduct experiments by assigning localities to bins randomly (randomly partitioned approach). Second, to provide spatially independent evaluation data, we implement geographically structured $k$-fold cross-validation as described above (taking background data from the full study region; geographically structured approach). Third, based on the principles of proper background selection, we modify the second approach in what we term the masked geographically structured approach. In this approach, we segregate the localities into bins in the same manner as the geographically structured one. However, for model calibration (and background sampling), we mask out environmental data from the entire area corresponding to the bin used for model evaluation in that iteration. Theory indicates that: 1) the first approach will not be able to detect any overfitting to sampling bias, leading to overestimates of
performance; 2) the second will be able to detect overfitting to bias (producing realistic estimates of performance), but will suffer from issues related to background sampling given the non-stationarity associated with the artificial bias introduced into the system; and 3) the third will be able to detect overfitting to bias, hence yielding realistic evaluations, but will not suffer from problems related to background sampling. In addition, estimates of performance for this approach assess model transferability, since neither the evaluation localities nor any background pixels from the area (bin) of the evaluation localities were used in model calibration. These experiments should shed light on issues regarding data-partitioning and background selection and resolve controversy differing conclusions reached regarding performance and transferability in recent studies (Peterson et al., 2007; Phillips, 2008; Anderson and Raza, 2010).

To demonstrate implementation of species-specific tuning with these data-partitioning approaches, we make models using Maxent (Phillips et al., 2006; Phillips and Dudík, 2008), which has performed well in recent studies and is in common use (Elith et al., 2006; Hernandez et al., 2006; Wisz et al., 2008). We conduct the tuning experiments in an area that we term the principal study region (where most of the localities of the species occur; see Materials and methods), with the aim of determining the optimal species-specific settings for reducing overfitting and hence producing a general and transferable model. To do so, we vary the regularization multiplier, a parameter that penalizes complex models; higher regularization values lead to simpler models (following Elith et al., 2010; see also Anderson and Gonzalez, unpublished data). For each value that we
use for the regularization multiplier, we run $k$ iterations (one for each bin used as evaluation data) for each of the three different data-partitioning approaches. For each iteration, we calculate threshold-independent and threshold-dependent measures of performance (see Materials and methods). We then take the average of the $k$ iterations for each combination of regularization multiplier and data-partitioning approach. Then, to assess the effect of regularization multiplier on model transferability, we apply (project) each model onto a second study region (hereafter, the projection region), where a distinct set of spatially independent localities exists for the study species. We gauge performance quantitatively in the projection region using the same measures as in the principal study region. We address the practicality of this approach by conducting research with a species that is well known and for which high-quality locality data exist. As a complement, future studies should compare the current results to those obtained with simulated species (Meynard and Quinn, 2007; Elith and Graham, 2009).
Materials and methods

Study species and occurrence records

*Heteromys anomalus*, the Caribbean spiny pocket mouse, lends itself well to this type of investigation. The distribution of this common, easily captured terrestrial rodent is well known and corresponds to specific vegetation types (Anderson, 2003b; Anderson and Gutiérrez, 2009). It generally inhabits extensive stands of mature or secondary deciduous and evergreen forests but also has been collected in areas under subsistence agriculture and in gallery forests or stands of woody vegetation in the *llanos* (savannas) of Venezuela. Typically, it ranges from sea level to ca. 1600 m in elevation. This natural history information facilitates interpretation of model predictions in geography. In addition, a large dataset exists, comprised of georeferenced localities based on museum specimens that were examined and verified by a specialist. In the course of the taxonomic revisions, the coordinates for those localities were determined with great care using a variety of detailed maps, gazetteers, publications by collectors, field notes, and correspondence with living collectors. Furthermore, the species provides two naturally segregated sets of occurrence records (Fig. 1): 1) the main distribution roughly linearly situated in a wide band across northern Colombia and Venezuela, as well as the islands of Trinidad, Tobago, and Isla Margarita; and 2) a geographically distinct (and possibly disjunct) documented distributional area in the Río Magdalena valley of Colombia to the south. The configuration of localities in the former allows for a convenient west-to-east partitioning into geographic bins in the principal study region.
Complementarily, the latter provides a test for assessing the spatial transferability of the models (projection region).

Georeferenced occurrence localities for *H. anomalus* came from Anderson (2003b) and Anderson and Gutiérrez (2009; in total, 270 unique localities). We then filtered localities to obtain the maximum number that were at least 10 km apart from one another (Anderson and Raza, 2010). This filtering reduces the degree of spatial autocorrelation in the occurrence localities and hence also should decrease any environmental bias due to the assumed uneven sampling by collectors (Reddy and Dávalos, 2003; Hortal et al., 2008; Loiselle et al., 2008; Boakes et al., 2010). This should lead not only to a more realistic estimate of the species’ niche, but also to more appropriate evaluation data (Veloz, 2009). Although the 10-km rule is arbitrary (see Pearson et al., 2007), given the topographic and environmental heterogeneity of this system, it should satisfy the above goals without dramatically reducing the number of localities available to calibrate and evaluate the model. Had we not filtered localities in this manner, reducing the effects of spatial autocorrelation, it is likely that any trends encountered in the results would have been even more pronounced. For each cluster of localities less than 10 km from each other, we determined the maximum number of localities that could be retained. When more than one optimal solution existed for a given cluster, we chose one randomly. After filtering, 124 unique localities remained in the principal study region and six in the projection region (Fig. 1; see below). Although few localities are available in the
projection region, we take advantage of them to conduct an additional independent test of transferability.

**Environmental data**

For the environmental data, we used 19 bioclimatic variables from WorldClim 1.4 (Hijmans et al., 2005; http://www.worldclim.org) at a resolution of 30 arc seconds (approximately 1 km\(^2\) near the equator). These variables are based on monthly precipitation and temperature data collected from weather stations. Hijmans et al. (2005) interpolated the monthly weather-station data spatially via a splining technique, taking elevation into account, and then derived the following 19 variables: annual mean temperature, mean diurnal range (temperature), isothermality, temperature seasonality, maximum temperature of the warmest month, minimum temperature of the coldest month, annual temperature range, mean temperature of the wettest quarter, mean temperature of the driest quarter, mean temperature of the warmest quarter, mean temperature of the coldest quarter, annual precipitation, precipitation of the wettest month, precipitation of the driest month, precipitation seasonality, precipitation of the wettest quarter, precipitation of the driest quarter, precipitation of the warmest quarter, and precipitation of the coldest quarter. Although some authors have examined the correlation structure among variables and removed some highly correlated variables (e.g., Veloz, 2009), other workers stress the use of prior knowledge and theoretical expectations in variable selection whenever possible (Mac Nally, 2000; Austin, 2002). These 19 bioclimatic variables have
predicted the distributions of other rodents successfully in this region (Anderson and Raza, 2010; Anderson and Gonzalez, unpublished data). Here, we use all of them to determine the behavior of Maxent with a set of variables that are likely to predict the distribution of this species and that show characteristics typical of those employed by many current modeling studies.

As described above, we defined two study regions. For the principal study region, which we use for the primary tuning experiment, we delimited a rectangle that surrounded the full extent of the known occurrences of the northern (coastal) distribution of the species. This area seems reasonable for approximating the assumptions of background selection by not including large regions that the species does not inhabit because of limitations to dispersal or because of biotic interactions (Anderson and Raza, 2010). Specifically, we used a rectangle whose borders were the nearest even half degree that was at least a half degree from the nearest locality; 7–13° N, 60–78° W). The projection region used for the independent test of transferability was delineated as the upper portion of the catchment of the Río Magdalena, where a geographically distinct set of localities exists. We used this region because the species is not known from any adjacent areas in the middle or upper drainage of the Río Cauca to the west, or the Orinocan and Amazonian catchments east of the crest of the Cordillera Oriental. Specifically, we drew a polygon from 6.5° N and continuing to the southern extreme of the basin, at approximately 1.5° N.

**Tuning experiments and data partitions**
In the tuning experiments, we made models using Maxent (Phillips et al., 2006). We used Maxent 3.2.1 and the logistic output option (Phillips and Dudík, 2008), running models via the command-line interface. To influence the level of model complexity, we calibrated models with different values for the regularization multiplier (see Introduction). In Maxent, regularization protects against overfitting by applying a penalty for each term included in the model and for higher weights given to each term, thus limiting model complexity (Phillips et al., 2006). This penalty occurs in the form of a $\beta$ regularization parameter specific to each feature class in the model (in a niche modeling context, a feature is an environmental variable or some function thereof; e.g., linear, quadratic, hinge, product; Phillips and Dudík, 2008). Current releases of Maxent implement a regularization multiplier, a user-specified number that is multiplied to the value of the $\beta$ parameter of each respective feature class, allowing the researcher to alter the overall level of regularization employed.

We varied the value of the regularization multiplier (0.25, 0.50, 1.00, 1.50, 2.00, 4.00, 6.00, 8.00, and 10.00) to assess what effect it had on model performance and transferability (Elith et al., 2010; Anderson and Gonzalez, unpublished data). At the sample sizes of localities used here in model calibration, Maxent allows use of all feature classes (linear, hinge, quadratic, product, threshold, and discrete); we used all except discrete, which only is relevant for categorical variables. Maxent addresses the issue of truncated response curves (see above), which often is relevant when transferring a model
across space or time, via the conservative assumption of clamping (Phillips et al., 2006; Anderson and Raza, 2010).

In all experiments, we divided the filtered localities falling in the principal study region into four bins of equal sample size (n = 31 in each bin; Figs. 1, 2). We chose four bins in order to have a substantial number of localities for evaluation without unduly reducing the number used for calibration. In contrast to the random approach, for the geographically structured and masked geographically structured approaches, we partitioned data spatially with four bins (each a rectangle) arranged longitudinally from west to east. Each had equal sample size, but they differed in area, together matching the extent of the full principal study region (longitudes for the bins: Bin A, 72.70–78.00° W; Bin B, 69.00–72.70° W; Bin C, 64.07–69.00° W; Bin D, 60.00–64.07° W). For the masked geographically structured approach, background selection of environmental data came only from the areas of the bins used in model calibration (“masking” out the area corresponding to the bin being withheld for evaluation by using a dummy variable; see Anderson and Raza, 2010).

**Model evaluation**

We assessed model performance using threshold-independent and threshold-dependent measures. As a threshold-independent assessment of overall model performance, we used the Area Under the Curve (AUC) of the Receiver Operating Characteristic plot. For presence-background evaluations, AUC quantifies the probability that the model correctly orders (ranks) a random
presence locality higher than a random background pixel (see Phillips et al., 2006 for use of AUC for presence-background assessment). AUC values calculated with presence–background evaluation data vary according to the proportion of the study region that is suitable for the species (Phillips et al., 2006). Hence, AUC values are not comparable among species or across study regions. However, they are appropriate for the present comparisons of the relative predictive ability of models produced with different settings but for the same species in the same study region (Lobo et al., 2008; Peterson et al., 2008). We obtained evaluation AUC from the Maxent in two different ways. Whenever possible, we extracted AUC from Maxent output files. However, Maxent does not automatically provide AUC values for projected models. Therefore, for the masked geographically structured approach and the projections to the Río Magdalena valley, we calculated it in a script using the AUC tool available in Maxent 3.2.17. For each data-partitioning approach, AUC was then averaged across the four iterations for each regularization multiplier in the principal study region. We then averaged AUC values for the projection study region in the same manner.

In addition to considering the respective evaluation AUC values in the principal and projection study regions, we calculated the difference between the calibration and evaluation AUCs in the principal study region. Calibration AUC and evaluation AUC values quantify performance on those respective datasets but do not directly measure overfitting. In contrast, by definition, the difference between the two quantifies overfitting. More specifically, in our data-partitioning
experiments, the magnitude of the difference between calibration and evaluation AUC quantifies the degree of overfitting to noise in the random approach, and overfitting to noise and/or bias in the geographically structured and masked geographically structured approaches (assuming stationarity of the species’ response across geography).

Complementarily, as a threshold-dependent evaluation, we measured the omission rate (the proportion of evaluation localities falling outside the prediction). We did so after applying two thresholds, the lowest presence threshold and the 10 percentile presence threshold. Under either thresholding rule, pixels with values equal to or higher than the threshold are considered suitable, whereas pixels with values below it are not, yielding a binary prediction of present vs. absent for the species. The lowest presence threshold (Pearson et al., 2007; = minimum training presence threshold of Maxent) is the lowest value of the prediction for any of the pixels that correspond to the calibration localities; hence, it indicates the least-suitable environmental conditions for which a locality was available in the calibration data set. Similarly, the 10 percentile presence threshold (= 10 percentile training omission threshold of Maxent software) sets as the threshold the value that excludes the 10 percent of the localities having the lowest predicted values. It constitutes a stricter (less permissive) criterion for converting a continuous prediction to a binary one, leading to a smaller geographic prediction for the species.

For an ideal model, we expect zero omission of evaluation localities using the lowest presence threshold and approximately 10 percent omission for the 10
percentile presence threshold. Omission rates that are higher than expected for a given threshold indicate overfitting, providing a second way of detecting overfitting to noise and/or bias. Because the lowest presence threshold is sensitive to the particular locality that is least suitable (which, in many cases, may have a substantially lower value than the next-least suitable one), it often may lead to an overly extensive prediction. In contrast, the 10 percentile presence threshold should not be nearly as sensitive to particular “outlier” localities and, hence, may provide more consistent results.

We averaged omission rate for the principal study region and then for the projection region as described above for AUC. Rather than employing omission rate to assess model significance (Anderson et al., 2002a), we use it (like AUC) to compare performance among model settings and data-partitioning approaches (see Anderson and Gonzalez, unpublished data). We interpret as optimal the regularization multiplier/s that reduced omission rate to the minimum or near-minimum level, minimized the difference between calibration and evaluation AUC, and still led to maximal or near maximal evaluation AUC. Similar to the situation for AUCs, we obtained omission rates from Maxent output files whenever possible but calculated them ourselves for the masked geographically structured approach and for the projection study region (Río Magdalena valley). Specifically, we extracted the value of the prediction at each evaluation locality using the GetVal tool in Maxent 3.2.17 and then determined the average omission rates in a spreadsheet.
We also evaluated model performance by qualitative visual examination of the resulting maps of the predicted potential distribution for the species, based on expert knowledge of the distribution of vegetation and habitat types that the species is known to inhabit. Although we examined maps of all predictions, for brevity we present only those for one bin and for selected regularization multipliers (0.25, 1.00, 2.00, 4.00, and 8.00). For each combination of data-partitioning approach and regularization multiplier, we observed: 1) whether the model showed signs of overfitting to calibration localities, 2) the strength of the prediction in the region of the excluded bin (not relevant to the randomly partitioned approach), 3) the overall discriminatory ability of the model, and 4) details of the predictions in particular regions where strong differences were apparent among regularization multipliers and/or data-partitioning approaches (e.g., the Sierra Nevada de Santa Marta, the Cordillera de Mérida, and the dry lowlands north and east of the Lago de Maracaibo; see Results). As signs of overfitting in geographic space, we searched for very small regions of high prediction lying close to calibration localities that do not correspond to recognized vegetation types. Similarly, we looked for the potential effects of the importance of single variables in the prediction (likely a form of overfitting)—sharp breaks in the strength of the prediction or alternating “ripples” in the prediction. The latter occurs when, for a given bioclimatic variable, the months included in particular quarters of the year jump or even alternate across the landscape. In addition, where relevant, we examined maps of clamping, to assess the degree to which it may have affected predictions.
Results

AUC

Average evaluation AUC (hereafter, AUC unless otherwise noted) in the principal study region remained relatively flat across the range of values for the regularization multiplier for all three approaches (Fig. 3a). However, each approach showed the highest AUC at the default regularization multiplier (1.00), and performance decreased slightly as the regularization multiplier was increased or decreased from the default. Across all values of the regularization multiplier, the geographically structured approach showed substantially lower AUC than did the random one. However, the masked geographically structured approach performed similarly to the random one.

In the projection region (Río Magdalena valley), AUC for all three approaches was highest for regularization multiplier values of 1.00 to 4.00 (Fig. 3b). For all three, AUC improved markedly from regularization multiplier values of 0.25 to 1.00, remained relatively flat until 4.00, then decreased sharply. In contrast to the results for the principal study region, the three data-partitioning approaches performed similarly here at low regularization values, but the masked geographically structured models (and to a lesser degree, the randomly partitioned ones) performed slightly worse than those from the geographically structured approach, at regularization multiplier values greater than 2.00.

All three approaches displayed similar trends regarding the difference between calibration and evaluation AUC in the principal study region. The difference was moderately high at low levels of the regularization multiplier but
rapidly decreased approaching the default setting (1.00) and leveled off at 4.00 (Fig. 4). Across all regularization multiplier values, the geographically structured approach displayed a notably higher difference than did the random approach. However, the difference for the masked geographically structured approach was extremely similar to that for the random one.

**Omission rate**

In the principal study region, average omission rate for the evaluation localities (hereafter omission rate) using the lowest presence threshold was very high for all three approaches at low regularization values but quickly dropped off for intermediate and high ones (Fig. 5a). The three curves were virtually flat above a regularization multiplier of 1.50, where rates were only slightly above the zero omission rate expected without overfitting (omission rate at regularization multiplier of 1.50: randomly partitioned, 0.065; geographically structured, 0.073; masked geographically structured, 0.032). The geographically structured approach displayed a higher average omission rate than the random approach did at regularization multiplier values of 0.25 to 1.00, but the two performed very similarly beyond that. The masked geographically structured approach performed similarly to the random one, but at regularization multipliers above 1.00, the omission rate was slightly lower for the former.

Using the 10 percentile presence threshold in the principal study region, all three approaches showed a pattern that is similar, but more pronounced than that for the lowest presence threshold (Fig. 5b). Extremely high omission rates
occurred at low regularization multipliers. Omission rates decreased markedly as the regularization multiplier increased; however, here they did not level off until a regularization multiplier of 4.00. Furthermore, as was expected, the lowest omission rates achieved were substantially higher than those for the lowest presence threshold for all approaches (Fig. 5a; at 4.00; randomly partitioned, 0.097; geographically structured, 0.250; masked geographically structured, 0.105). Across all regularization multipliers, the geographically structured approach showed a higher omission rate than did the random one. However, the omission rates for the masked geographically structured approach were almost identical to those of the random one. At regularization multipliers of 4.00 and above, the omission rates of the random and masked geographically structured approaches were only slightly higher than expected without overfitting (0.10, or 10%), but that for the geographically structured approach was substantially higher.

In the projection region (Río Magdalena valley), the omission rate using the lowest presence threshold showed similar patterns to those in the principal study region but was substantially higher at low regularization values (Fig. 5c). For all three approaches, omission rates were extremely high at regularization multiplier 0.25 and decreased precipitously then moderately from 0.25 to 4.00, with the curves either flat or decreasing slightly after that. Minimal differences existed among the three approaches, which all led to omission rates slightly to moderately above expected (zero) at regularization multipliers at and above 4.00
(randomly partitioned, 0.042; geographically structured, 0.125; masked geographically structured, 0.083).

Using the 10 percentile presence threshold in the projection region, trends were similar to those above for all three approaches (Fig. 5d), but omission rates were much higher than for the two relevant comparisons: the lowest presence threshold in the projection region (Fig. 5c) and the 10 percentile presence threshold in the principal study region (Fig. 5b.). Omission rates were extremely high at low regularization multipliers but decreased as regularization value was increased, leveling off at values of 4.00 to 6.00. As for this thresholding rule in the principal study region, the omission rate at those high regularization values was higher for the geographically structured approach (0.417) than for the other two (0.292, both). However, all three approaches showed omission rates far above that expected without overfitting (0.10).

**Qualitative assessments**

In geographic space, the predictions differed dramatically among regularization values and, to a lesser degree, among data-partitioning approaches. Although trends were similar for all bins, we present and interpret only those for Bin B and for selected regularization multipliers (Fig. 6; 0.25, 1.00, 2.00, 4.00, and 8.00). For all approaches, the signs of overfitting decreased markedly with increased levels of regularization, but the very highest regularization values led to models that failed to capture important aspects of the species’ potential distribution (based on expert knowledge). Models made with
the lowest regularization multiplier (0.25) suffered from extreme overfitting, with the strongest predictions restricted to areas near calibration localities. At the default regularization multiplier (1.00), overfitting was substantially lower, but many sharp borders were evident in the predictions, likely due to the effect of single variables. Models made with regularization multipliers 2.00 and 4.00 were rather similar. At 2.00, fewer sharp borders occurred than at regularization multiplier 1.00, and the areas strongly predicted for the species generally corresponded to vegetation types where the species is known to occur. Generally good discrimination between suitable and unsuitable environments was found at high elevations (Sierra Nevada de Santa Marta and Cordillera de Mérida). Models for regularization multiplier 4.00 were very similar to those for 2.00, but they lost some discriminatory ability in the higher-elevation areas mentioned above. Although the models made using regularization multiplier 8.00 appear grossly similar to those at 4.00, they have lost substantial discrimination in general and do not reflect the species’ tolerances accurately in highland areas, showing dramatic overprediction at the highest elevations.

The masked geographically structured approach led to more realistic predictions in geography than did the other data-partitioning approaches (Fig. 6). At most regularization multipliers, the prediction in the area corresponding to Bin B (the evaluation bin for the predictions described here; Figs. 1, 2) was weaker for the geographically structured approach than for the random one. The geographically structured model made using the lowest regularization multiplier (0.25) gave very weak prediction for most of the evaluation localities. For that
approach, the other examined regularization multipliers shared the common pattern of a much weaker prediction in piedmont areas of the Cordillera de Mérida (in the area of Bin B). Interestingly, however, regularization multipliers 1.00, 2.00, and 4.00 indicated a much stronger prediction in piedmont areas of the Sierra Nevada de Santa Marta (outside the area of Bin B). In comparison with these first two approaches, the masked geographically structured approach showed a notably stronger prediction in the area of Bin B. This difference was always apparent in comparisons with the geographically structured approach (where it often was very strong) and existed in comparisons with the random approach for regularization multipliers up to and including 2.00. The model for regularization multiplier 0.25 indicated an extremely strong prediction for the area of Bin B but overall showed less overfitting than did either of the corresponding models for the other two approaches. Additionally, the model for regularization multiplier 1.00 (and to a lesser degree that for 2.00) indicated a stronger prediction than those for the other two approaches in the dry areas along the Caribbean coast to the north and east of the Lago de Maracaibo. For models made with regularization multipliers 2.00 and 4.00, overall better discrimination was apparent for this approach than for the random one. For all regularization multipliers, clamping was minimal in the masked geographically structured approach—appreciable only in small areas along the Caribbean coast and at the southern end of the Lago de Maracaibo (not shown).
Discussion

Interpretation of tuning experiments

AUC

The threshold-independent evaluations using AUC indicate differences in performance among regularization multipliers and among approaches for data-partitioning, but patterns differed between the principal and projection study regions. In the principal study region, varying the regularization multiplier has only moderate effect on AUC (with a weak peak around the default value), but marked differences exist among the three approaches (Fig. 3a). Although the randomly partitioned models have higher AUCs than the geographically structured ones, this difference probably derives from both artifactual and real causes. As mentioned above, evaluation localities are not independent from the calibration data in the randomly partitioned approach; hence, its estimate of performance is likely inflated (Araújo et al., 2005b; Veloz, 2009), in part because any overfitting to bias cannot be detected. In contrast, the data for the geographically structured approach indeed have more bias (because of additional, strong bias inserted by us), and the corresponding evaluation can detect overfitting to it. Therefore, we suspect that some of the observed difference in performance between the two derives from an overinflated AUC for the random approach, but that the rest of the difference correctly reflects an overfitting to the artificial bias inserted by segregating our data spatially in the geographically structured approach (see also Anderson and Raza, 2010). In contrast, the masked geographically structured models enjoy estimates of
performance as high as those for the randomly partitioned ones, despite being able to detect any overfitting to bias. The difference in performance between the masked geographically structured approach and the geographically structured one (which takes background samples from the entire principal study region) emphasizes the importance of selecting calibration regions that match the assumptions of modeling (Anderson and Raza, 2010). Including background data from a region that includes evaluation localities but not calibration ones (in the geographically structured approach) provides a false negative signal that interferes with successful modeling of a species’ environmental requirements, here decreasing model performance dramatically.

In contrast, in the projection study region (Río Magdalena valley), varying the regularization multiplier has a strong effect on AUC, but little difference exists among the three approaches at the regularization values that had the highest AUC values (1.00 to 2.00; Fig. 3b). The peak in AUC corresponds to the default regularization value and slightly to moderately higher ones. Clearly, both extremely low and exceedingly high regularization lead to poor performance, indicating low transferability, under each partitioning approach. Perhaps most importantly, the apparently superior performance of the random approach in the principal study region disappears here in the projection study region, which provides an independent test in which measures of performance should not be artifactually inflated for the random approach.

The difference between calibration and evaluation AUC in the principal study region detects strong overfitting at low regularization values (Fig. 4).
Because the evaluation AUCs are almost flat (Fig. 3a), the observed differences (which quantify overfitting) derive from trends in the calibration AUC (not shown). The geographically partitioned approach shows much higher overfitting than do the other two. This difference between it and the other two approaches matches almost perfectly the magnitude of the corresponding difference observed in evaluation AUC in the principal study region (Fig. 3a). Hence, the three approaches vary little in calibration AUC. As in the interpretations of evaluation AUC, the higher overfitting indicated here for the geographically structured models probably corresponds both to their ability to detect overfitting to bias and to the higher level of bias present in this approach. Again, the estimates of overfitting for the random approach are likely to be somewhat depressed (yet overly optimistic), since this approach cannot detect any overfitting to bias. However, the masked geographically structured models show performance nearly identical to that of the randomly partitioned ones here, and this low level of overfitting is realistic (i.e., not a possible artifact, as in the random approach). These interpretations clarifying the differences among the approaches (considering the effects of overfitting to bias) leave a common pattern intact: all three curves show a striking decline from the lowest regularization multipliers to a value of 2.00, which we attribute to a decline in overfitting to noise.

Omission rate

The threshold-dependent evaluations of omission rate also show differences in performance among regularization multipliers and among data-
partitioning approaches, and patterns once more vary between the principal and projection study regions. Here, patterns in omission mirror those found for the difference between calibration and evaluation AUC, showing results rather different from those for evaluation AUC in either the principal or projection regions. All four combinations of evaluation region and thresholding rule illustrate a marked decrease in omission rate as the regularization multiplier increases from the lowest value to moderate ones. However, the respective curves level off at different regularization multipliers, and the best (lowest) omission rate achieved in a given analysis varies between thresholding rules and between the two regions.

The four combinations reach their respective lowest (or essentially the lowest) omission rates at different regularization multipliers. The use of the lowest presence threshold in the principal study region achieves nearly zero omission at regularization multipliers of 1.50 and above (Fig. 5a). The corresponding value is higher for the 10 percentile presence threshold (Fig. 5b). Both thresholding rules show an additional increase in the lowest omission rate achieved in the projection region (Río Magdalena valley), not leveling off until a very high regularization multiplier (4.00 to 6.00; Figs. 5c, d).

Curiously, whereas the three approaches show virtually identical performance using the lowest presence threshold, the 10 percentile presence threshold indicates a notable difference among approaches, with similar trends in both regions (Figs. 5b, c). Using the 10 percentile presence threshold, the geographically structured approach shows higher omission rates than the other
two, at least at the high regularization multipliers that lead to the lowest omission rates. Just as for the lower evaluation AUC and the higher difference between calibration and evaluation AUC (see above), we ascribe this pattern to both a higher level of bias in this approach (inserted by the experimental design itself) and the ability to detect overfitting to it. The fact that the lowest presence threshold does not detect a similar difference among approaches may somehow be related to its being very sensitive to the particular locality that is least suitable.

Whereas the lowest omission rates achieved are acceptably low in the principal study region (at least for two of the approaches; Fig. 5a, b), conclusions for the projection region differ between the two thresholding rules (Fig. 5c, d). In the principal study region and using the lowest presence threshold, the masked geographically structured approach attains the expected zero omission rate or a value only slightly higher than it (Fig. 5a). The other two approaches yield omission rates that are slightly higher but still acceptable. Using the 10 percentile presence threshold, both the random and masked geographically structured approaches achieve excellent omission rates (near the expected 10%), but the higher rate mentioned above for the geographically structured approach is unacceptably high (almost 30%; Fig. 5b).

For the projection region (Río Magdalena valley), the lowest presence threshold achieves omission rates that are only moderately higher than expected, but those for the 10 percentile presence threshold are very high (and clearly unacceptable). Specifically, the omission rates achieved here by the lowest presence threshold are acceptable at and above a regularization multiplier of
4.00, depending on the data-partitioning approach (Fig. 5c). In contrast, the 10 percentile presence threshold never achieves acceptably low omission rates, showing particularly high ones for the geographically structured approach (Fig. 5d). As in the differences among data-partitioning approaches (see above), the discrepancy here between the two thresholding rules may again indicate that the lowest presence threshold does not provide a realistic indicator of model performance and likely overestimates performance (at least with the current data, especially given the large sample size of occurrence localities).

The unacceptably high omission rate (for the 10 percentile presence threshold) in the projection region has several possible explanations, related to a lack of transferability and/or stationarity. We explore these alternatives but note that the few evaluation localities in the projection region preclude firm conclusions at present. First, the high omission rate could be due to an overfit model that transfers poorly. However, the masked geographically structured approach itself evaluates transferability among subregions (bins) of the principal study region, and the evaluations for this approach indicated acceptable omission (and hence, high transferability) of models made with intermediate and high regularization multipliers. The other possible explanations concern a lack of stationarity, which also lowers transferability. Notably, clamping in the Río Magdalena valley was minimal at and near the evaluation localities there (strong clamping only occurred at the far southern extreme of the basin and at high elevations at the northern extreme of the eastern slope of the Cordillera Central—far from known localities. This indicates that any differences in the
available environments did not strongly affect the model transferal. Additionally, a difference in biotic context between the principal and projection study regions is possible (e.g., competitive release in the projection region, leading to increased omission rates there for a model made in the other region). However, in the present system, the distributions of congeneric species do not appear to correspond to such a situation; indeed, geographic analyses actually suggest competitive release for *H. anomalus* in the principal study region (Anderson et al., 2002b). Alternatively, a lack of stationarity may be due to a difference in inherited niche characteristics between the two regions, a possibility that we cannot address with the data at hand. Firm conclusions regarding these issues must await future research.

**Visual interpretations in geography**

As judged by visual interpretations of the predictions in geography, for all approaches, low regularization multipliers produce problematic levels of overfitting, intermediate ones yield satisfactory predictions, and the highest multipliers lead to underfit models that show unrealistic predictions in some regions (Fig. 6). Overfitting is very strong at the lowest regularization value and even to some degree at the default one (1.00) but is not apparent (or only weakly so) at and above regularization multiplier 2.00. Models made with regularization multipliers 2.00 and 4.00 seem highly appropriate and correspond most closely to the distribution of the vegetation types that the species is known to inhabit (Anderson, 2003b; Anderson and Gutiérrez, 2009). However, at regularization
multiplier values of 8.00 and 10.00, the models lose the ability to discriminate some aspects of the species’ potential distribution, notably overpredicting in the highest elevations.

As predicted, the three data-partitioning approaches differ strongly in their predictions with regard to the area corresponding to the bin used for evaluation, with the masked geographically structured approach leading to the most realistic predictions overall (Fig. 6). The geographically structured approach consistently underpredicts the area corresponding to the bin used for evaluation (indicating overfitting to bias), but the masked geographically structured approach rectifies this shortcoming, at least at intermediate to high regularization multipliers. Oddly, the latter approach leads to overly strong predictions in the area of the evaluation bin at low and even some intermediate regularization multipliers. Apparently, this overprediction does not derive from issues related to clamping, which was minimal for all regularization multipliers. We interpret that the especially low prediction for piedmont areas in the Cordillera de Mérida in the geographically structured approach is due to a general overfitting to bias. In contrast, the strong prediction for such regions in the Sierra Nevada de Santa Marta (in which the species has been especially frequently collected; Anderson, 2003b) in both the geographically structured approach and the masked geographically structured approach (at least at intermediate regularization values) probably derives from the inclusion of many more localities in those calibration datasets relative to the random approach.
Conclusions

Taking all estimates of performance into account, we interpret that optimal performance for models of *H. anomalus* in this region with these environmental variables corresponds to regularization multipliers higher than the default (1.00). Although a slight peak occurs in AUC at the default regularization value in the principal study region, the highest values generally correspond to regularization multipliers between 1.00 and 4.00 in the projection study region. More importantly, however, all other measures of performance (both in the principal and projection regions) indicate much better performance at regularization multipliers slightly to substantially higher than default. Specifically, in the principal study region, based on the difference between calibration and evaluation AUC as well on omission rates using both thresholding rules, regularization multipliers as high as 2.00 to 4.00 are necessary to reduce overfitting to acceptable levels. Qualitative assessments of the geographic predictions reiterate this conclusion. Omission rates in the projection region suggest slightly higher regularization values as optimal, but some of that increase may derive from a lack of stationarity (see above). In sum, regularization multipliers of 2.00 to 4.00 achieve acceptable performance regarding omission while still maintaining peak or near-peak AUC values. Although AUC values and omission rates do not worsen with regularization multipliers above 4.00 (except for AUC in the projection region), qualitative assessments of models in geography show a decline in model quality and overall discriminatory ability.
These results echo the findings of two recent studies, where regularization settings higher than default are required to achieve optimal model performance, especially when transferability is required (Elith et al., 2010; Anderson and Gonzalez, unpublished data). However, in the geographically structured approach, increasing the regularization multiplier is insufficient to counteract the extremely strong bias in the localities used for model calibration (artificially inserted in that data-partitioning approach). The masked geographically structured approach rectifies the problem of artificial bias inserted by the researcher, but the results suggest that if extremely strong environmental bias exists in the locality data for a given species, integration of information regarding sampling effort is necessary (Phillips et al., 2009).

**Recommendations**

The current results lead us to recommendations regarding the use of model tuning (or “smoothing”) to identify optimal model complexity for a given species and dataset. For Maxent, ideally both the regularization multiplier employed and the feature classes considered (e.g., linear, quadratic, hinge, product) should be subjected to tuning experiments. Regarding regularization, we suggest that the optimal settings should be the lowest regularization multiplier that meets the following criteria. Foremost, the settings must lead to acceptable omission rates near those expected for the threshold employed, and minimize the difference between calibration and evaluation AUC—indicating low overfitting. Secondarily, of the settings that achieve the first criterion, preference
should be given to those that maximize evaluation AUC (show the highest discriminatory ability). Finally, if multiple settings result from these optimality criteria, the lowest of the co-optimal regularization multipliers should be chosen (that least likely to underfit the model). Future research also should determine if varying the regularization multiplier (which preserves the relative strengths of the $\beta$ regularization parameter across feature classes) is sufficient to achieve optimal regularization values (i.e., rather than tuning $\beta$ individually for each feature class; Anderson and Gonzalez, unpublished data). Examination of the species’ response curves to individual variables likely will provide additional information regarding overfitting (Elith et al., 2010). Finally, qualitative visual assessment of model predictions in geography can complement (and hopefully corroborate) this quantitative perspective. Here, with a well-known species, qualitative assessments were feasible and informative (matching the conclusions of the quantitative evaluations), but with more poorly known species for which this is not possible, researchers can rely on quantitative evaluations to guide them in selecting optimal model settings. Similar efforts to achieve optimal model complexity should be undertaken with other modeling techniques (e.g., GAM/GLM, boosted regression trees, GARP, etc.).

Even if transferability per se is not required in the application at hand, masked geographically structured data-partitioning (such as this implementation of $k$-fold cross-validation) holds great promise in the context of model evaluation and tuning for Maxent and any other presence-background technique. Under most circumstances, it should provide realistic measures of model performance
(in contrast to the random approach; Veloz, 2009) without violating assumptions of modeling (in contrast to the geographically structured approach where background data are selected from the entire study region; Anderson and Raza, 2010). While we used a simple west-to-east partitioning strategy with only four bins, both the geographic arrangement and the number of bins should be tailored to the project at hand. The number of localities within each bin should be roughly equal, as here, because the number of localities used to calibrate the model will impact its output (Phillips and Dudík, 2008; Wisz et al., 2008). Overall, for species with many localities, the use of a greater number of bins may be feasible. Additionally, the geographic configuration of bins likely will vary according the shape of the species' known distribution (rather than a simple rectangle, as here). Ideally, the environmental conditions available and biotic contexts should not differ among bins.

Under this overall strategy, two paradigms exist to produce a final “best” model for a species. One possibility is to determine optimal settings as here and then calibrate a final model using those settings and all localities (i.e., from all bins, not withholding any for evaluation). However, the settings that are best for one sample size may not be the same as those that are optimal for the larger sample that uses all localities. Alternatively, the $k$ models created with the optimal settings (using the masked geographically structured approach) could be combined into a composite prediction. Such a composite would provide both 1) a consensus average prediction for the species in each map pixel and 2) an estimate of the variability of the prediction for each pixel, an important added
benefit. The drawback is that not all localities are used in model calibration, ignoring potentially useful information and increasing the possibility of problems related to any lack of stationarity among subregions of the species’ distribution.

To reach general conclusions regarding model tuning, comprehensive experiments like the present study are necessary, but with multiple species, varied numbers of localities, and different numbers and types of environmental variables. Such research also should examine the effects of spatial autocorrelation in the localities (e.g., with different levels of filtering of calibration localities) as well as of the level of correlation among environmental variables (Elith et al., 2010). As an ultimate test with empirical species, the approach suggested here should be compared with that of removing the effects of sampling bias when it can be quantified directly or estimated using a suitable target group (Anderson, 2003a; Phillips et al., 2009). Additionally, similar research should be undertaken with simulated species (Meynard and Quinn, 2007; Elith and Graham, 2009). This overall research agenda may allow for a complex set of rules for estimating the optimal settings for Maxent based on variables such as the sample size of localities, level of spatial autocorrelation in the localities, sampling bias (if quantifiable), and the number, kind, and level of correlation among the environmental variables. Such experiments using truly independent (e.g., geographically structured) evaluation data might eventually lead to replacement of the general guidelines produced by Phillips and Dudík (2008). In the meantime, species-specific tuning of model settings holds substantial promise for improving ecological niche models whenever time and
resources permit. To facilitate tuning, scripts such as those used here can be utilized to run models in batches and to extract values for quantitative measures of model performance (i.e., AUC and omission rates). In addition, automation of other aspects of the process would allow much-greater use of the approaches espoused here.
Acknowledgments

This research was made possible by funding from the U. S. National Science Foundation (NSF DEB-0717357, to Robert P. Anderson) and the Office of the Dean of Science and Office of the Provost (The City College of The City University of New York). Funds to present preliminary results from this research were generously provided by the International Biogeography Society (Student Travel Award, to AR). Innumerable thanks go to Robert P. Anderson for his guidance and mentorship throughout the course of this study and throughout my tenure at City College of New York; his patience and sagacity were critical both to the completion of this work and my development as a researcher. Darla M. Thomas assisted with data preparation and preliminary modeling, and her sunny disposition made much of the tedious early work a pleasure. Robert A. Boria, Ana C. Carnaval, Eliécer E. Gutiérrez, Ali Raza, Jhanine L. Rivera, Mariya Shcheglovitova, and Mariano Soley-G. provided insightful comments and suggestions that greatly improved the manuscript. Steven J. Phillips provided instruction in running Maxent from the command line and answered several queries about other Maxent functionalities. Finally, I thank Amy C. Berkov and David J. Lohman for serving on my committee and for their constructive critique of the manuscript.
References


**Figure 1.** Filtered localities of the Caribbean spiny pocket mouse, *Heteromys anomalus* and study regions in northwestern South America used in this study. Note 1) the principal study region in the north showing four geographic bins (A, B, C, and D) along the Caribbean coast; and 2) the projection region in the Río Magdalena valley. Shaded areas correspond to elevations above 1000 m. Localities that appear to fall on the border between bins are located in either one or the other, but not both.
Figure 2. Example of data-partitioning approaches for tuning experiments of Maxent models of the Caribbean spiny pocket mouse *Heteromys anomalus*. Black circles represent localities used for model calibration and white ones denote localities used for evaluation. Shaded areas shown correspond to one environmental variable (annual mean temperature) for the areas used for background sampling.
Randomly partitioned

Geographically structured

Masked geographically structured
Figure 3. Results of threshold-independent evaluations (AUC) for tuning experiments of Maxent models of the Caribbean spiny pocket mouse, Heteromys anomalus, for all data-partitioning approaches in the principal study region (a) and projection region (b). For all three approaches, evaluation AUC was averaged across the $k$ iterations of each value of the regularization multiplier. In the principal study region, AUC remained relatively flat across the range of regularization multiplier values for all three approaches; note, however, the noticeable difference in performance between the geographically structured approach and the other two. In contrast, in the projection region, AUC for all three approaches peaked at intermediate regularization values, with little difference in performance among the three approaches.
a. Principal study region

b. Projection region
Figure 4. Plot of the average difference between calibration and evaluation AUC in the principal study region for tuning experiments of Maxent models of the Caribbean spiny pocket mouse, *Heteromys anomalus*, for all data-partitioning approaches. This difference, which quantifies overfitting, was moderately high at low levels of regularization but rapidly decreased approaching the default setting and leveled off at regularization multiplier 4.00. Across all regularization multipliers, the geographically structured approach displayed a notably higher difference than did the random or masked geographically structured approach.
The graph shows the relationship between the regularization multiplier and the calibration AUC – evaluation AUC for different data structures: randomly partitioned, geographically structured, and masked geographically structured. As the regularization multiplier increases, the calibration AUC – evaluation AUC decreases, indicating improved calibration and evaluation performance.
Figure 5. Results of threshold-dependent evaluations (omission rate) for tuning experiments of Maxent models of the Caribbean spiny pocket mouse *Heteromys anomalus* for all data-partitioning approaches in the principal study region (a = lowest presence threshold; b = 10 percentile presence threshold) and the projection region (c = lowest presence threshold; d = 10 percentile presence threshold). All four combinations of region and thresholding rule illustrated a marked decrease in omission rate as the regularization multiplier increased from the lowest value to moderate ones. However, the respective curves level off at different levels of regularization, and the best (lowest) omission rate achieved in a given analysis varied between thresholding rules and between the two regions. In both regions and across all values of regularization multiplier, the omission rate was lower using the lowest presence threshold (a, c) than using the 10 percentile presence threshold (b, d). However, omission rates using either rule were higher in the projection region than in the principal study region.
a. Principal study region, lowest presence threshold

![Graph showing average omission rate against regularization multiplier for different data structures.]

b. Principal study region, 10 percentile presence threshold

![Graph showing average omission rate against regularization multiplier for different data structures.]
c. Projection region, lowest presence threshold

![Graph showing the average omission rate against the regularization multiplier for different observation windows.](image)

d. Projection region, 10 percentile presence threshold

![Graph showing the average omission rate against the regularization multiplier for different observation windows.](image)
**Figure 6.** Maxent models of the potential geographic distribution of the Caribbean spiny pocket mouse *Heteromys anomalus* in the principal study region for all the data partitioning approaches and selected regularization multipliers. The predictions show a suitability gradient from low (0, blue) to high (1, red). Squares correspond to calibration (white) and evaluation (purple) localities (evaluation localities not shown for masked geographically structured approach). Although we only present models for which Bin B constituted the evaluation data, trends were similar for the other bins. For all approaches, the effects of overfitting decreased markedly with increased regularization, and the highest regularization values led to models that failed to capture important aspects of the species’ potential distribution. For the masked geographically structured approach evaluation localities (although not shown) are the same as those for the geographically structured approach.
Masked geographically structured

Geographically structured

Randomly partitioned

Regularization multiplier

0.25 1.00 2.00 4.00 8.00