Short communication

Spatial filtering to reduce sampling bias can improve the performance of ecological niche models

CrossMark

Robert A. Boria [a,*], Link E. Olson [b], Steven M. Goodman [c,d], Robert P. Anderson [a,e,f]

[a] *Department of Biology, City College of the City University of New York, New York, NY 10031, USA*
[b] *University of Alaska Museum, Fairbanks, AK 99708, USA*
[c] *Field Museum of Natural History, Chicago, IL 60605, USA*
[d] *Association Vahatra, BP 3972, Antananarivo 101, Madagascar*
[e] *Graduate Center, City University of New York, New York, NY 10016, USA*
[f] *Division of Vertebrate Zoology (Mammalogy), American Museum of Natural History, New York, NY 10024, USA*

ARTICLE INFO

ABSTRACT

This study employs spatial filtering of occurrence data with the aim of reducing overfitting to sampling bias in ecological niche models (ENMs). Sampling bias in geographic space leads to localities that may also be biased in environmental space. If so, the model can overfit to those biases. As a preliminary test addressing this issue, we used Maxent, bioclimatic variables, and occurrence localities of a broadly distributed Malagasy tenrec, *Microgale cowani* (Tenrecidae: Oryzorictinae). We modeled the abiotically suitable area of this species using three distinct datasets: unfiltered, spatially filtered, and rarefied unfiltered localities. To quantify overfitting and model performance, we calculated evaluation AUC, the difference between calibration and evaluation AUC ($=AUC_{diff}$), and omission rates. Models made with the filtered dataset showed lower overfitting and better performance than the other two suites of models, having lower omission rates and AUC_{diff}, and a higher $AUC_{evaluation}$. Additionally, the rarefied unfiltered dataset performed better than the unfiltered one for three evaluation metrics, likely because the larger one reinforced the biases. These results indicate that spatial filtering of occurrence localities may allow biogeographers to produce better models.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Ecological niche models (ENMs) are a correlative approach aiming to approximate the abiotically suitable area of a species by comparing environmental conditions at localities where the species occurs with the overall conditions available in the study region (see Peterson et al., 2011; Anderson, 2012 for terminology). The increased prevalence of online databases of occurrence localities and climatic variables has resulted in an increase in the production of ENMs (Hijmans et al., 2005; Kozak et al., 2008). Although correlative ENMs are used widely in the fields of ecology, evolution, and conservation biology, their mainstream acceptance has outpaced methodological research and refinement.

Here, we study one area needing methodological improvement: the effect of sampling bias. Frequently, researchers sample easily

accessible areas (i.e., near major roads or towns), leading to geographic clusters of localities (Hijmans et al., 2000; Kadmon et al., 2004; Reddy and Dávalos, 2003). These sampling biases artificially increase spatial auto-correlation of the localities. Such a situation can cause the model to overfit to environmental biases that correspond to these influences in geographic space. Overfitting occurs when a model fits too tightly to calibration data, limiting the model's ability to predict independent evaluation data. Eliminating artifactual clusters of localities is also important for model evaluation, since calibration localities that are next to evaluation localities lead to inflated values of performance (Hijmans, 2012; Veloz, 2009).

In this study, we aim to reduce the effect of sampling bias by spatially filtering the occurrence dataset, which should reduce the degree of overfitting in the model. Ideally, when information quantifying sampling effort exists (e.g., via a target group), it can be used in model calibration to correct for sampling bias (Anderson, 2012; Phillips et al., 2009). However, researchers frequently do not have access to such information. In contrast, the method applied here can be employed generally. Several studies have used filtering (=thinning) techniques (Anderson and Raza, 2010; Carroll, 2010; Pearson et al., 2007; Veloz, 2009) to reduce the effects of sampling

* Corresponding author at: 160 Convent Avenue, Marshak Science Building, Room J-526, New York, NY 10031, USA. Tel.: +1 212 650 8424.
E-mail addresses: rboria00@citymail.cuny.edu,
robertboria@gmail.com (R.A. Boria).

biases, but we know of none that have explicitly tested whether this method improves the performance of ENMs (but see Varela et al., 2013 for an implementation with a virtual species). If it does, an ENM made with the filtered dataset should show lower overfitting and higher performance in predicting independent evaluation data.

2. Methods and materials

2.1. Occurrence and environmental data

Madagascar is home to four endemic radiations of extant terrestrial mammals, including nesomyine rodents, lemurs, euplerid carnivorans, and tenrecs. The latter shows considerable morphological variation and forms an extraordinary adaptive radiation (Olson and Goodman, 2003), with the most taxonomically diverse genus being the shrew tenrecs (*Microgale* spp.; 22 currently recognized extant species; Goodman et al., 2006; Olson, 2013; Olson et al., 2009; Soarimalala and Goodman, 2011). Perhaps the most common, widespread, and well-documented species, Cowan's shrew tenrec (*Microgale cowani*) is found throughout what remains of Madagascar's humid forests at elevations ranging from 530 to 2500 m (Soarimalala and Goodman, 2011). This swath spans several different vegetational zones, including forests ranging from lowland to upper montane, as well as ericoid alpine formations above the forest line. This species appears to be a generalist among shrew tenrecs and accounts for over one-fifth of *Microgale* specimens in European and North American museums (Olson, unpub.). Because its range and habitat requirements are relatively well known, *M. cowani* represents a suitable species for the current study.

Occurrence localities were compiled from field collections and associated notes, examination of museum specimens, and literature (Fig. 1, appendix). The environmental data were obtained from WorldClim.org (Hijmans et al., 2005; at 30 sec. resolution). These 19 bioclimatic variables employed reflect aspects of temperature and precipitation and have been used successfully for niche models of small non-volant montane mammals (e.g., Jezkova et al., 2009; Davis et al., 2007). We delimited a custom study region for each model, specifically by drawing a rectangle around localities and adding a 0.5° buffer (Anderson and Raza, 2010; Barve et al., 2011; see Fig. 1.).

2.2. Experimental design

As a first exploration, we built models using Maxent version 3.3.2k. Maxent is a presence-background algorithm that compares occurrence localities with a sample of background pixels to create a prediction of suitability (Phillips et al., 2006; Phillips and Dudík, 2008). Maxent has performed well in comparison with other techniques and is commonly used (Elith et al., 2006; Wisz et al., 2008) but sensitive to sampling biases (Anderson and Gonzalez, 2011; Phillips et al., 2009). In addition to sampling bias, two other issues can affect overfitting in niche models: correlations among environmental variables and the level of model complexity. To simplify the current experiment, we held those factors constant. Specifically, we used all 19 bioclimatic variables and employed default Maxent settings for the given sample size: feature class (linear, quadratic, and hinge) and regularization multiplier value (1). We note, however, that Maxent employs regularization to reduce complexity; because of this, not all variables are necessarily included in the final model (Phillips and Dudík, 2008).
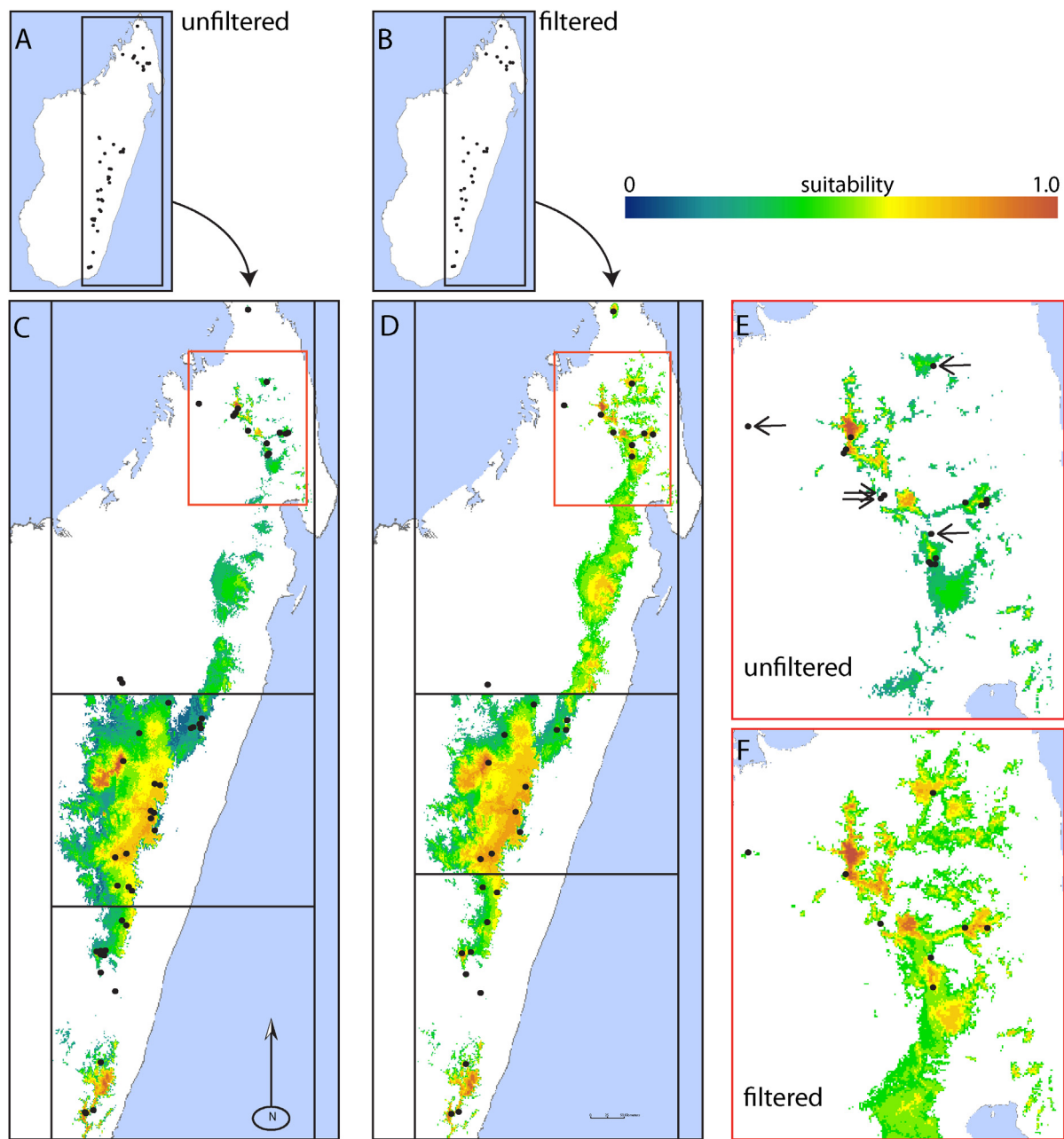
For filtering, we randomly removed localities that were within 10 km of one another, keeping the most localities possible. The 10 km distance was chosen based on the high spatial heterogeneity of the mountains in Madagascar, and the same distance has been used in previous studies in mountainous areas with high

geographical heterogeneity (Pearson et al., 2007; Anderson and Raza, 2010). This distance was not chosen to approximate the species' dispersal capabilities, but rather to reduce the inherent geographic biases associated with collection data. There were 57 unique localities before filtering and 31 unique localities after filtering (see Fig. 1). We used the Geographic Distance Matrix Generator version 1.2.3 to calculate the geographic distance between each pair of localities (Ersts, 2012). For each cluster of localities less than 10 km apart, we determined the maximum number of localities that could be retained. When more than one co-optimal solution existed for a given cluster, we selected one randomly. To test for the expected effect of reducing sampling bias versus simply the effect of sample size, we also randomly rarefied the unfiltered dataset to match the number of localities of the filtered dataset. Hence, we used three different datasets for modeling: unfiltered, filtered, and rarefied unfiltered. To explore the possibility that the spatial filter used here removed localities with novel environmental conditions, we plotted the values of annual mean temperature and annual mean precipitation at each locality.

An overfit model has an overly complex relationship between the occurrence localities of a species and associated environmental variables (Peterson et al., 2011). To quantify overfitting as well as general model performance, we implemented a variation of k-fold cross-validation. To provide strong tests, we divided the localities *geographically* into k = three bins (see Fig. 1). Each bin was constructed to contain approximately the same number of localities but occupy different portions of geography (Radosavljevic and Anderson, 2013). This allowed the models to be evaluated on spatially segregated (spatially independent) evaluation data, avoiding the inflation of evaluation metrics due to spatial autocorrelation between calibration and evaluation datasets (Hijmans, 2012; Veloz, 2009). Such evaluations also are necessary for evaluating model transfer across space or time (e.g., for climate change studies; Anderson, 2013). In each iteration, the models were calibrated using $k - 1$ bins and evaluated on the withheld bin (Fielding and Bell, 1997; Peterson et al., 2011). This was done until all bins were used once for evaluation (i.e., three iterations in total). By using custom study regions for each iteration, Maxent sampled background data for the environmental variables from only the regions corresponding to the bins used during calibration (following Phillips, 2008; Radosavljevic and Anderson, 2013). These methods allowed quantification of overfitting and performance after transfer (Peterson et al., 2011; Araújo and Rahbek, 2006; Bahn and McGill, 2013). The model from each iteration was then projected to the full study region to allow for evaluation and visualization.

We evaluated overall model performance via threshold-independent and threshold-dependent measures that assess various aspects of performance and overfitting. The threshold-independent metrics derive from the Area Under the Curve (AUC) of the Receiver Operating Characteristic plot, a rank-based measure of overall discriminatory ability of the model. Accordingly, the AUC calculated on evaluation localities ($AUC_{evaluation}$) constituted our measure of overall model performance. The other threshold-independent measure was AUC_{diff}: $AUC_{calibration}$ minus $AUC_{evaluation}$. The smaller the difference between the two, the lesser the overfitting present in the model (Warren and Seifert, 2011). Because comparisons between AUCs calculated using presence-background data are only valid when study regions are identical, we calculated AUCs over the entire study region. For each iteration of each treatment, we obtained AUC_{diff} and $AUC_{evaluation}$, and then averaged the values across the three geographic bins.

Complementarily, we employed two threshold-dependent measures: omission rates based on two threshold rules (10% calibration omission rate and lowest presence threshold, LPT = 0% calibration omission rate; Pearson et al., 2007; =minimum training

Fig. 1. Maps showing the effects of geographic filtering on ecological niche models of the shrew tenrec *Microgale cowani* in Madagascar. Results are shown for unfiltered localities (A and C) and geographically filtered localities (B and D). Geographically structured *k*-fold cross-validation (see localities indicated within *k* = 3 rectangles) allowed evaluation of performance of spatially segregated (spatially independent) localities. Models calibrated with two of the three bins were then projected to the geographic region corresponding to the withheld bin (three projected models shown together in C and D). The right panel provides details for a region in northern Madagascar for each dataset (E, unfiltered; F, filtered), with arrows indicating localities omitted by the respective prediction. Increasingly warm colors indicate successively stronger predictions. For each withheld region, areas in white indicate those pixels with values below the LTP threshold determined based on calibration data. Note that, overall, the filtered dataset led to broader predictions, omitted fewer evaluation localities, and predicted the evaluation localities more strongly.

presence threshold of Maxent software). Omission rates are the proportion of evaluation localities that fall outside of the model once converted to a binary prediction. Omission rates provide information regarding both discriminatory ability and overfitting, evaluated at particular thresholds. In general, lower omission indicates higher performance (better discrimination between suitable versus unsuitable areas). In addition, overfit models have omission rates higher than theoretical expectations (Radosavljevic and Anderson, 2013; Shcheglovitova and Anderson, 2013). The lowest presence threshold sets the threshold at the lowest value of the prediction for any pixel that contains a calibration locality (Anderson and Gonzalez, 2011) and has an expected omission rate of zero for evaluation localities. Similarly, the 10% calibration omission rate

rule sets the threshold at a value that excludes the 10% of calibration localities with lowest prediction (Anderson and Gonzalez, 2011) and has an expected omission rate of 0.10. We obtained the two threshold-dependent measures and averaged their values as for AUC.

3. Results

3.1. Evaluations

The models made with the filtered dataset were superior to those from the other datasets for all metrics. Regarding the threshold-independent measures, the filtered datasets led to the

Table 1

Summary statistics for performance and overfitting in experiments assessing the effects of geographic filtering in ecological niche models of the shrew tenrec *Microgale cowani* in Madagascar. The three datasets of localities (unfiltered, filtered, and rarified unfiltered) were each subjected to spatially independent evaluations using geographically structured *k*-fold cross-validation. Results are averaged for one measure of overall performance ($AUC_{evaluation}$) and three measures of overfitting: AUC_{diff} ($AUC_{calibration} - AUC_{evaluation}$), and average omission rates (lowest presence threshold and 10% calibration omission threshold). Note that the filtered dataset showed the lowest value for each measure of overfitting, as well as the highest value for the measure of overall performance.

Dataset	AUC		Omission rate	
	AUC_{diff}	$AUC_{evaluation}$	Lowest presence threshold	10% calibration omission threshold
Unfiltered	0.057	0.903	0.193	0.263
Rarefied unfiltered	0.043	0.912	0.164	0.382
Filtered	0.028	0.914	0.100	0.194

highest average $AUC_{evaluation}$ (highest overall performance) and the lowest average AUC_{diff} (lowest overfitting; Table 1; Supplementary Table 1). For the threshold-dependent measures, models made with the filtered datasets showed the lowest average omission rate for each of the threshold rules (lowest overfitting and highest discriminatory ability at those thresholds; Table 1; Supplementary Table 2). Despite the observed pattern for $AUC_{evaluation}$, average differences among the three suites of models were small in magnitude for that measure. In contrast, the magnitude of differences in the other three measures was substantial. The models made with the filtered dataset showed the lowest divergence from the expected values for both omission rates, indicating the lowest overfitting. Importantly, the models made with the rarefied unfiltered dataset outperformed the unfiltered dataset for three of the four evaluation metrics (AUC_{diff}, $AUC_{evaluation}$, and LTP omission rate), but for all measures they still performed more poorly than the filtered datasets. Notably, spatial filtering did not remove any localities holding novel environmental conditions for the examined variables (Supplementary Fig. 2).

Qualitative evaluations in geographic space illustrate spatial patterns that help interpret the quantitative results (Fig. 1.). Inspection of the three respective projection regions (each corresponding to the respective bin that was withheld during calibration) using the LPT threshold indicated the extent of predicted areas and the localities that were omitted for the unfiltered and the filtered datasets. All analyses predicted moderate to high suitability for the species throughout several extensive areas, in accordance with known natural history information for this widespread species occurring across a broad elevational range and in several mesic vegetational zones. However, the filtered dataset showed a generally broader prediction. The map for the rarefied unfiltered dataset (Supplementary Fig. 1) was similar to that for the unfiltered dataset. The models made with the filtered dataset predicted more evaluation localities than those for the other two treatments (Fig. 1 and Table 1).

4. Discussion

4.1. Comparisons among datasets

We show that for *M. cowani*, spatial filtering of occurrence datasets led to a reduction of overfitting and an increase in performance. The effects of filtering were not merely due to sample size, because the filtered dataset also outperformed the rarefied unfiltered dataset. The fact that the rarefied dataset outperformed the unfiltered dataset could be due to the larger dataset reinforcing the biases (as in Anderson and Gonzalez, 2011). Although the three datasets led to similar levels of overall performance based on $AUC_{evaluation}$, the filtered dataset did yield the highest value for that measure. Hence, all measures indicated that the filtered dataset provided the best prediction of spatially independent occurrence localities of this species.

Whereas sampling bias generally goes unaccounted for in ecological niche modeling, the current results support the value of

spatial filtering and point to several directions for future research. When available, data quantifying sampling biases should be used to correct for those biases (e.g., via a target group; Phillips et al., 2009). However, for the many cases when such data are not available, spatial filtering of datasets may improve model performance. The spatial distance used here for filtering was chosen based on the heterogeneity of the study system. In future research, we recommend that researchers conduct filtering experiments to determine the optimal geographic distance for the organism(s) and system(s) while retaining the species-specific niche signal (Anderson, 2012). We predict that mountainous regions will require a spatial filter that is smaller compared with regions having more homogenous environments. Furthermore, in addition to filtering in geographic space, researchers should consider the possibility of filtering localities in environmental space (Varela et al., 2013; see Supplementary Fig. 2). Finally, future comprehensive filtering experiments should be undertaken while varying the environmental datasets employed (e.g., different numbers of variables, and considering the correlations among them) and model complexity (e.g., in Maxent by changing feature classes and regularization), ideally with both simulated and real species.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ecolmodel.2013.12.012.

References

Anderson, R.P., 2012. Harnessing the world's biodiversity data: promise and peril in ecological niche modeling of species distributions. Annals of the New York Academy of Sciences 1260, 66–80.

Anderson, R.P., 2013. A framework for using niche models to estimate impacts of climate change on species distributions. Annals of the New York Academy of Sciences 1297, 8–28.

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

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

Araújo, M.B., Rahbek, C., 2006. How does climate change affect biodiversity. Science 313, 1396–1397.

Bahn, V., McGill, B.J., 2013. Testing the predictive performance of distribution models. Oikos 122, 321–331.

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S.P., Peterson, A.T., Soberón, J., Villalobos, F., 2011. The crucial role of the accessible area in ecological niche modeling and species distribution modeling. Ecological Modelling 222, 1810–1819.

Carroll, C., 2010. Role of climatic niche models in focal-species-based conservation planning: assessing potential effects of climate change on Northern Spotted Owl in the Pacific Northwest, USA. Biological Conservation 143, 1432–1437.

Davis, E.D., Koo, M.S., Conroy, C., Patton, J.L., Moritz, C., 2007. The California Hotspots Project: identifying regions of rapid diversification of mammals. Molecular Ecology 17, 120–138.

Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S., Wisz, M.S., Zimmerman, N.E., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29, 129–151.

Ersts, P.J., 2012. [Internet] Geographic Distance Matrix Generator (version 1.2.3). American Museum of Natural History. Center for Biodiversity and Conservation, Available from http://biodiversityinformatics.amnh.org/open_source/gdmg (accessed 21.05.12).

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

Goodman, S.M., Raxworthy, C.J., Maminirina, C.P., Olson, L.E., 2006. A new species of shrew tenrec (*Microgale jobihely*) from northern Madagascar. Journal of Zoology 270, 384–398.

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

Hijmans, R.J., Garrett, K.A., Huaman, Z., Zhang, D.P., Schreuder, M., Bonierbale, M., 2000. Assessing the geographic representativeness of genebank collections: the case of the Bolivian wild potatoes. Conservation Biology 14, 1755–1765.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25, 1965–1978.

Jezkova, T., Jaeger, J.R., Marshall, Z.L., Riddle, B.R., 2009. Pleistocene impacts on the phylogeography of the Desert Pocket Mouse (*Chaetodipus penicillatus*). Journal of Mammalogy 90, 306–320.

Kadmon, R., Farber, O., Danin, A., 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. Ecological Applications 14, 401–413.

Kozak, K.H., Graham, C.H., Wiens, J.J., 2008. Integrating GIS-based environmental data into evolutionary biology. Trends in Ecology and Evolution 23, 141–148.

Olson, L.E., 2013. Tenrecs. Current Biology 23, R5–R8.

Olson, L.E., Goodman, S.M., 2003. Phylogeny and biogeography of tenrecs. In: Goodman, S.M., Benstead, J. (Eds.), The Natural History of Madagascar. The University of Chicago Press, Chicago, pp. 1235–1242.

Olson, L.E., Rakotomalala, Z., Hildebrandt, K.B.P., Lanier, H.C., Raxworthy, C.J., Goodman, S.M., 2009. Phylogeography of *Microgale brevicaudata* (Tenrecidae) and description of a new species from western Madagascar. Journal of Mammalogy 90, 1095–1110.

Pearson, R.G., Raxworthy, C., Nakamura, M., Peterson, A.T., 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. Journal of Biogeography 34, 102–117.

Peterson, A.T., Soberón, J., Pearson, R.G., Anderson, R.P., Martínez-Meyer, E., Nakamura, M., Araújo, M.B., 2011. Ecological Niches and Geographic Distributions. Princeton University Press, Princeton.

Phillips, S.J., 2008. Transferability, sample selection bias and background data inpresence-only modelling: a response to Peterson et al (2007). Ecography 31, 272–278.

Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., Ferrier, S., 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications 19, 181–197.

Phillips, S.J., Dudík, M., 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31, 161–175.

Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190, 231–259.

Radosavljevic, A., Anderson, R.P., 2013. Making better Maxent models of species distributions: complexity, overfitting, and evaluation. Journal of Biogeography, Epub ahead of print.

Reddy, S., Dávalos, L.M., 2003. Geographical sampling bias and its implications for conservation priorities in Africa. Journal of Biogeography 30, 1719–1727.

Shcheglovitova, M., Anderson, R.P., 2013. Estimating optimal complexity for ecological niche models: a jackknife approach for species with small sample sizes. Ecological Modelling 269, 9–17.

Soarimalala, V., Goodman, S.M., 2011. Les petits mammifères de Madagascar. Association Vahatra, Antananarivo.

Varela, S., Anderson, R.P., García-Valdés, R., Fernández-González, F., 2013. Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. Ecography (in press).

Veloz, S.D., 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. Journal of Biogeography 36, 2290–2299.

Warren, D.L., Seifert, S.N., 2011. Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. Ecological Applications 21, 335–342.

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A., 2008. Effects of sample size on the performance of species distribution models. Diversity and Distributions 14, 763–773.