Ecological Modelling 275 (2014) 73–77

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Short communication

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

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

Article history:
Received 27 August 2013
Received in revised form
10 December 2013
Accepted 12 December 2013
Available online 14 January 2014

Keywords:
Ecological niche models
Madagascar
Overfitting
Sampling bias
Spatial filter
Tenrecidae

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).

0304-3800/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ecolmodel.2013.12.012

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 3.3.2*k*. 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

Download English Version:

https://daneshyari.com/en/article/4376003

Download Persian Version:

https://daneshyari.com/article/4376003

Daneshyari.com