Ecological Modelling 386 (2018) 83–88

Contents lists available at ScienceDirect

Ecological Modelling

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

Short communication

The effect of large sample sizes on ecological niche models: Analysis using a North American rodent, *Peromyscus maniculatus*

Robert A. Boria*, Jessica L. Blois

University of California, Merced, Merced, CA, 95343, USA

ARTICLE INFO

Keywords:
Occurrence records
Ecological niche model
Maxent
Peromyscus maniculatus
Sample size
Species distribution model

ABSTRACT

Correlative ecological niche models (ENMs) aim to approximate the environmentally suitable areas for a species. Recently, studies have explored the minimum number of occurrence records needed to implement ENMs; however, cosmopolitan species with many occurrence records have their own challenges and the effects of larger sample sizes on ENM performance have yet to be determined. To address this issue, we focused on a New World rodent, *Peromyscus maniculatus*. We obtained locality data from GBIF (13,199 unique records), and spatially filtered the localities. We then modeled suitable area for the species using Maxent, two different environmental datasets (at different spatial resolutions) and different numbers of occurrence records (with 25 replicates per sample size). We evaluated the models with k-fold cross-validation, AUC, and two omission rates. Further, we calculated the variability among predictions within and between datasets to indicate variation in geography. Generally, the AUC and omission rate both decreased as sample size increased. Lastly, as sample size increased, similarities in geography increased within and between datasets. For *P. maniculatus*, we get similar performing models, both in terms of geographic predictions and evaluation statistics, with as few as 10%–20% of the maximum number of localities for each environmental dataset. Using a large number of occurrence records may not be necessary for ENMs, and in fact may hinder model performance.

1. Introduction

Ecological niche models (ENMs; oft termed Species Distribution Models) are correlative models that approximate the environmentally suitable areas for a species. They do so by comparing the overall environmental conditions available to the species with the conditions at localities where the species occurs (Peterson et al., 2011). Because ENMs can be projected across time and space they are frequently used in a wide range of ecological and evolutionary studies. However, despite their relevance to many fields, several methodological issues (such as, sample bias; correlation of variables; model evaluation) hinder their implementation in many systems (Anderson, 2012; Merow et al., 2013). Here, we study the effects of large numbers of occurrence records on ENMs.

Several studies have determined the minimum number of occurrence records needed to implement ENMs for different algorithms (Papeş and Gaubert, 2007; Proosdij et al., 2016; Stockwell and Peterson, 2002; Wisz et al., 2008), revealing that as few as five localities can produce biologically meaningful models (Galante et al., 2018; Pearson et al., 2007; Shcheglovitova and Anderson, 2013). However, cosmopolitan species with many occurrence records present their own challenges and the effects of larger sample sizes on models have yet to be determined. These species typically have large geographic extents encompassing heterogeneous environments, which could potentially pose problems when trying to estimate the environmental suitability of areas. Using too few localities may not capture the species' entire niche; however, using too many occurrence records may not increase model accuracy and could potentially hinder model performance (i.e., by limiting the amount of available background localities and reducing discriminatory ability; Stockwell and Peterson, 2002; VanDerWal et al., 2009). Further, as sample size increases so do associated issues of sampling bias and georeferencing errors, which could potentially lead to overfit models (Bloom et al., 2018; Boria et al., 2014).

In this study, we aim to determine how large sample sizes affect ENM evaluations and predictions in geography. Specifically, using a widely distributed species (*Peromyscus maniculatus*), we calibrated and evaluated our models using a spatial partition method, two different environmental datasets (at different spatial resolutions) – worldclim (Hijmans et al., 2005) and Community Climate Simulation Model 3 (CCSM3; Lorenz et al., 2016) – and an increasing number of localities. Further, we evaluated the predictions in geographic space to determine differences between the different datasets when they are projected

* Corresponding author at: 5200 North Lake Road, Science and Engineering Building 1, Merced, CA, 95343, USA.
E-mail address: robertboria@gmail.com (R.A. Boria).

https://doi.org/10.1016/j.ecolmodel.2018.08.013
Received 5 June 2018; Received in revised form 21 August 2018; Accepted 22 August 2018

across space. We explored if using smaller datasets would produce similar models to larger datasets, without noticeable improvements in model performance and in predictions. The results suggest that as sample size increases there isn't necessarily a noticeable increase in model performance. When modeling species with a large number of occurrence records, it may not be necessary to use all localities for ENMs and could potentially affect model performance negatively.

2. Methods

2.1. Input data

We conducted analyses with a New World rodent, *Peromyscus maniculatus* (North American deer mouse), which is indigenous to and distributed widely across North America (Shorter et al., 2012). This species can be found in every terrestrial ecosystem, and is the most diverse and widespread species of deer mouse (Bedford and Hoekstra, 2015; Dewey and Dawson, 2001). This species provides a good system because there are currently more than 100,000 occurrence records for *P. maniculatus* in the Global Biodiversity Information Facility (GBIF). We downloaded all GBIF records that contained coordinates, a preserved specimen, and didn't have any geospatial errors, using the R (v. 3.4.1, R Development Core Team, 2016) package *rgbif* (Chamberlain, 2017). We then removed all duplicate records, and all localities that were placed in the ocean and outside the known geographic range (based on the NatureServe range estimate; Patterson et al., 2007). These steps led to a final set of 13,199 unique *P. maniculatus* localities. Despite efforts to improve data quality, we acknowledge errors associated with GBIF data (e.g., Beck et al., 2014); however, the main goal of this study is the effects of large number of occurrence records on ENMs and not determine *P. maniculatus* environmental suitability.

Generally, easily accessible roads are sampled more frequently, generating issues with sampling bias (Hijmans et al., 2000; Kadmon et al., 2004; Reddy and Dávalos, 2003). Additionally, spatial biases within GBIF datasets have been shown to negatively affect ENM performance (Beck et al., 2014). To reduce sampling bias, we spatially filtered the occurrence dataset to ensure that no two localities were within a pre-determined distance of one another (while retaining the most localities possible) using the *spThin* package in R (Aiello-Lammens et al., 2015). This method effectively reduces artificially induced spatial auto-correlation (Boria et al., 2014; Fourcade et al., 2014; Kramer-schadt et al., 2013). Because of the difference in resolution among our climate predictor datasets (*see below*), we used a different distance for each climate dataset, based on their grid cell size: 25 km for the worldclim dataset and 75 km for CCSM3. The worldclim dataset was reduced to 3334 localities after applying the spatial filter, and the CCSM3 dataset was condensed to 990 occurrence records. From these 'full' datasets, we then randomly sampled different numbers of occurrence records, with 25 replicates for each sample size. The worldclim sample sizes were: 25; 50; 100; 500; 1000; 1500; 2000; 2500; 3000. For CCSM3 we used: 25; 50; 75; 100; 150; 200; 300; 400; 500; 600; 700; 800. These were the occurrence datasets used in the modeling exercises described below.

We used two different climate simulations that include climate inferences for the present day: worldclim (Hijmans et al., 2005) and Community Climate Simulation Model 3 (CCSM3; Liu et al., 2009, downscaled by Lorenz et al., 2016). These two simulations emphasize different tradeoffs: worldclim variables are downscaled to 1 km × 1 km grid cells for the entire world for present day, as well as 7000 and 21,000 years ago, and thus have better spatial resolution. CCSM3 variables are downscaled to 0.5 × 0.5 degrees (~50 km × 50 km) grid cells for North America from 21,000 years ago to the present day at 500-year intervals, and thus have better temporal resolution. Although we do not consider paleoclimates in this paper (all ENMs are based only on contemporary climate inferences), our ultimate research goal is to infer *P. maniculatus* climate suitability and hindcast the ENMs to the past, which motivated our inclusion of these two climate models. The worldclim dataset has 19 bioclimatic variables that reflect aspects of temperature and precipitation (Hijmans et al., 2005). The worldclim variables are known to be correlated; however, MaxEnt is a machine-learning algorithm that uses regularization to reduce complexity (especially regarding correlated variables), and thus not all variables are necessarily included in the final model (Elith et al., 2011; Phillips and Dudík, 2008). The CCSM3 environmental variables consist of 27 variables that also reflect features of temperature and precipitation (Lorenz et al., 2016). However, for the CCSM3 variables we followed the procedure of Maguire et al. (2016) and only used the six least correlated variables over the last 21,000 years (minimum precipitation of the driest quarter, maximum temperature of the warmest quarter, mean yearly potential evapotranspiration, maximum precipitation of the wettest quarter, mean yearly water deficit index and mean yearly actual evapotranspiration).

To approximate modeling assumptions regarding dispersal and biotic interactions more closely, we delimited a custom study region for the full dataset (13,199 localities; Appendix 1 in Supplementary file), specifically by drawing a minimum convex polygon around the localities and adding a 0.5° buffer (Anderson and Raza, 2010; Barve et al., 2011). Background localities for calibration (default number of 10,000) were taken from within the delimited study region only.

2.2. Ecological niche modeling

We used a machine learning approach, Maxent (v 3.3.3k; Phillips et al., 2006; Phillips and Dudík, 2008), to generate the ENMs. This method is a presence-background technique and has performed well in comparison with other techniques (Elith et al., 2006; Wisz et al., 2008; but see Fitzpatrick et al., 2013; Royle et al., 2012). To simplify the current work, we employed the default settings (feature class and regularization) in Maxent for each sample size.

We calibrated and evaluated the ENMs using a spatial partition approach using the R package *ENMeval* (Muscarella et al., 2014). The geographic range was divided into four quadrants $k = 4$, with approximately equal number of occurrence records. We then built models using a jackknife approach ($k - 1$) and evaluated models within the unused partition (Boria et al., 2014; Radosavljevic and Anderson, 2014). Maxent sampled background data for the environmental variables from only the regions corresponding to the quadrants used during calibration (Phillips et al., 2008). We did this a total of four times for each dataset so that each partition was used for evaluation once; we calculated evaluation statistics (see below) for each quadrant and averaged across the four iterations.

We evaluated the models using threshold-independent and -dependent methods. The threshold-independent measure, Area Under the Curve (AUC) of the Receiver Operating Characteristic plot, is a rank-based measure of discriminatory ability of the model. We calculated these two ways: 1) using all localities for each dataset (according to the chosen sample size) projected across the entire study region (Full model AUC), 2) calculating AUC for each evaluation quadrant and averaging the four iterations (Mean AUC). The threshold-dependent measures were based on different omission rate thresholding rules: 1) Minimum Training Presence (MTP) and 2) 10% calibration omission rate. Omission rate is the proportion of evaluation localities that are not correctly predicted as present, and measures model overfitting. Overfit models have omission rates higher than theoretical expectations. The MTP sets the threshold at the smallest value of the prediction for any grid cell that contains a calibration locality and has an expected omission rate of zero for evaluation localities and is a more conservative measure of model fitness. Similarly, the 10% calibration omission rate rule sets the threshold at a value that excludes the 10% of calibration localities with lowest prediction and has an expected omission rate of 0.10 (Pearson et al., 2007). We averaged the evaluation statistics across all 25 replicates for each sample size and calculated