BIEN Range Methods Description

Cory Merow

June 5, 2017

This document provides a brief overview of the methods used to develop range models for the BIEN3 database so that users can judge their adequacy for their own applications. Occurrence records were cleaned to resolve taxonomic naming issues and remove records were latitude/longitude was not available or could not be verified. Records that were cultivated or nonnative were removed, though native species lists were not available throughout the New World so this filtration was imperfect. Environmental covariates were obtained from WorldClim at 10 km resolution (Hijmans *et al.* , 2005). Predictors included mean annual temperature, mean diurnal temperature range, annual precipitation, precipitation seasonality, precipitation in warmest quarter/ (precipitation in warmest quarter + precipitation in coldest quarter), and five spatial eigenvectors. The spatial eigenvectors corresponded to large scale regional differences and primarily served to limit predictions far from presence locations in geographic space (Diniz-Filho & Bini, 2005). Only one occurrence record per cell (in cases of multiple records) was used for model building.

Different range estimation methods were used depending upon the sample size of (unique) presence locations. A species with a single record were assigned a range that included only the $100km^2$ cell where it was found. Ranges for species with 2-3 records were built with bounding boxes (area bounded by the minimum and maximum latitude and longitude of all occurrences). Ranges for species with 4-9 records were built with convex hulls (the minimum-fitting polygon that can be drawn to encompass all species occurrences). For species with ¿9 records, we built species distribution models using the Maxent algorithm (Phillips *et al.* , 2006). Maxent model building generally followed the recommendations outlined in Merow *et al.* (2013) and recommendations in Merow *et al.* (2014) for building relatively less complex models. Model settings were chosen to balance overfitting (under estimating range sizes) with underfitting (excessively smooth models that over predict range size). Only linear, quadratic, and product features were used and regularization was set at the default value Maxent's continuous predictions were converted to binary presence/absence predictions by choosing a threshold based on the 75^{th} percentile of the cumulative output (based on analyses validated with 700 species for which expert maps were available).

Automating model building for 90,000 species is not without flaws and some caveats should be recognized. Notably, sample size remains small for the vast majority of species, hence many ranges are estimated using some somewhat coarse methods (i.e. not from species distribution models). It is impossible to automatically detect all problematic, outlying, or nonnatural occurrence records and those that remain may influence range predictions. Given our attempts to avoid overfitting, the species distribution models are more likely to underfit spatial distribution patterns and consequently may predict ranges larger that those realized for some species. That is, the models may predict suitable habitat in locations that are inaccessible to the species (but in in similar environmental conditions to where they occur) or predict suitable habitat slightly beyond realized range edges due to fitting relatively smoothed response curves. To offset this, cells where presence was predicted by Maxent further than 1000km from any presence record were removed from the range. Correction has not been made to account for variation in sampling effort or detection probability. Like any range map, our predictions represent hypotheses about spatial occurrence patterns. In spite of these caveats, predictions for the vast majority of species are reliable and are well-suited for macroecological analyses.

Our range modeling efforts are a dynamic enterprise and we are constantly exploring ways to improve predictions, leading to periodic updates in our database. Planned updates include choosing optimal models settings tuned specifically for each species, accounting for sampling variation, and improving occurrence data cleaning methods. We will employ version control to maintain accessibility of all past versions as updates are released.

References

Diniz-Filho, JAF, & Bini, Luis Mauricio. 2005. Modelling geographical patterns in species richness using eigenvector based spatial filters. *Global Ecology and Biogeography*, **14**(2), 177–185.

Hijmans, Robert J, Cameron, Susan E, Parra, Juan L, Jones, Peter G, & Jarvis, Andy. 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**(15), 1965–1978.

Merow, Cory, Smith, Matthew J, & Silander, John A. 2013. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, **36**, 1–12.

Merow, Cory, Smith, Mathew J, Edwards, Jr, Thomas C, Guisan, Antoine, Mcmahon, Sean M, Normand, Signe, Thuiller, Wilfried, Wüest, Rafael O, Zimmermann, Niklaus E, & Elith, Jane. 2014. What do we gain from simplicity versus complexity in species distribution models? *Ecography*, **37**, 1267–1281.

Phillips, Steven J, Anderson, Robert P, & Schapire, Robert E. 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.