# Continuously-updated, bias-corrected species distribution modeling of natural history specimen databases: `tropicosMassModel` using `enmSdmBayes`

Adam B. Smith | Missouri Botanical Garden | adam.smith@mobot.org | 2020-04

## Introduction

Natural history specimen records represent the fundamental basis for inferring the distribution of Earth's several million species. However, despite centuries of collection, our knowledge of species' true distributions remains woefully inadequate, leading to the so-called "Wallacean shortfall." Understanding the true distribution of species is critical for conserving Earth's biodiversity and for answering long-standing questions in ecology. Fortunately, species distribution modeling can serve as a stand-in for thorough sampling.

To date, however, applications of species distribution models have wrestled with the uneven sampling inherent in biodiversity specimen collection database. We know, for example, that collections are much more likely to occur near populated places, universities, herbaria, and natural history museums, and along roads. Although techniques exist to correct for these issues, without a true estimate of bias, all bias-correction methods are hypotheses that cannot be confirmed.

The `tropicosMassModel` system is intended to rectify these issues. Specifically, `tropicosMassModel` models the distribution of specimen data using a modeling system derived from occupancy-detection methodology. To fit within this framework, specimens from the focal species background are used as an index of collection effort, and counties (or equivalent) are used as sample sites. Our work demonstrates that the breadth of the background used for effort has little effect on the model outcome, so long as there are sufficient records and geographic coverage to serve as an index of collection behavior.

The output of the `tropicosMassModel` system is a set of files, one per species that has been modeled, containing an image (Figure 1) of maps of the predicted distribution of the species, a shapefile with model predictions that can be downloaded and used in other software, and a file with information on the data and the model. These can be served by a publicly-facing web portal to the database. In our case, the system creates models of species for TROPICOS, the Missouri Botanical Garden's primary plant specimen database. Our hope is that this installation serves as a guide and inspiration for others to follow.

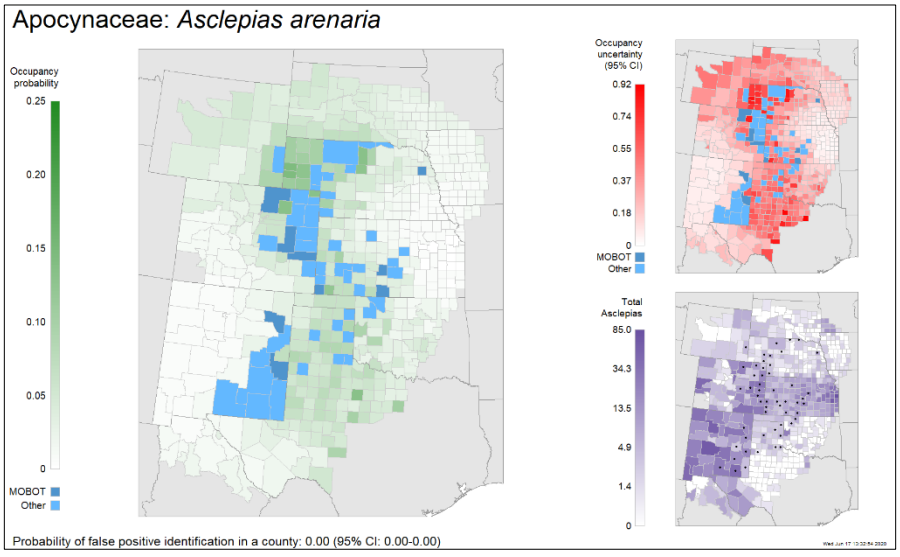


**Figure 1**. Example of output from `tropicosMassModel`. The image can be served to users on a biodiversity database. The map on the left shows known occurrences and the probability of occurrence in areas where the species has not been found. On the upper right is uncertainty in the estimate of the probability of occurrence. The bottom right shows the density of all specimens collections. Note the uneven density of specimens between states. The model corrects for this and for variation in collection effort within states/provinces.

The latest version of `tropicosMassModel` (and this document) can be obtained from https://github.com/adamlilith/tropicosMassModeling.

## Installation

Our installation of `tropicosMassModel` is particular to our use-case, but it is general enough to be broadly applicable. A modern computer with sufficient RAM is required (we use a 32-GB machine).

1. Install R (free, from https://www.r-project.org/).
2. Install RTools (https://cran.r-project.org/bin/windows/Rtools/). Please note that you may need to follow directions on that page for the installation to work.
3. Copy the entire `tropicosMassModel` repository from GitHub (https://github.com/adamlilith/tropicosMassModeling) into a folder on the machine. You may rename the folder if you wish.
4. Inside the folder open the script `tropicosMassModel.r` in a text editor. You will need to change the lines (near the top):

   ```
   source('C:/Ecology/Drive/R/tropicosMassModeling/tropicosMassModel.r')
   ```

   and

   ```
   # set working directory
   setwd('C:/Ecology/Drive/R/tropicosMassModeling')
   ```

   so it points to the directory in which the repository has been copied.

5. Start R and copy your version of the "source" line:

   ```
   source('C:/Ecology/Drive/R/tropicosMassModeling/tropicosMassModel.r')
   ```

   into R and hit Enter. R may download some other packages, and you may have to specify a server from which to obtain them (any is fine). If this is the first time you have run `tropicosMassModel` on this computer (and in this folder), the script will take a few hours to create files that are used for modeling. Once created (and if R is stopped and restarted), these will not have to be created again.

   The script should start processing each species, determining if it has sufficient records and background effort for modeling, and make models for those that do. Model output is stored in the folder `./models`. Not all models are statistically robust. Those that are robust are stored in the subfolder `./models/ok`, while those that are not are stored in `./models/notOk`. We strongly suggest not using the models in the "not OK" directory (saving output from these models can be turned off).

## Settings

The script `tropicosMassModel` has several settings that can be used to control the modeling behavior. All of these are near the top of the file under `settings`:

```
niter
nburnin
nchains
thin
```

These control the behavior of the Monte Carlo Bayesian model called by the script. In general they should not be changed, but smaller numbers (especially for `niter` and `nburnin`) can be used to speed modeling (i.e., for troubleshooting). In that case, we suggest using `niter = 100, nburnin = 10, thin = 1`, and `nchains = 2`. These values will almost always produce insufficient models (and if they do should still not be used).

`minUnconverged` Maximum proportion of unconverted nodes in model for the model to be considered "sufficient"

`minNumDupsToModel` Minimum number of duplicate records in a county necessary to generate a model

`minNumEffortsToModel` Minimum number of records of other species that must be in the study region to model the focal species

`minPropCountiesWithEffortToModel` Minimum proportion of counties in the study region with any effort necessary to model the species

`minEmptyCountiesPerState` Minimum number of counties to include state in modeling if it has no detections

`expand` Index of the size of the modeling region created for each species. Specifically, this is a multiplier of the largest distance from centroid of all occurrences to the farthest occupied county. The modeling region is a all counties touched by a buffer around the occurrences with a size equivalent to this distance times the multiplier. Numbers must be > 0 and are typically within ~0.2 and 0.5.

`minOccCounties` Minimum number of counties a species must appear in to model

`minNumRecords` Minimum number of records (across/within counties) to model

## Data

Eventually we will set up `tropicosMassModel` to obtain data directly from TROPICOS. At the moment, due to bandwidth constrains, the TROPICOS API has a limit of 500 records, which is insufficient for modeling most species. Hence, the script currently draws specimens from the BIEN database. This will be rectified in the future as new hardware allows increasing bandwidth for TROPICOS.

## Under the hood

For species with a sufficient number of records and background samples in its focal region, `tropicosMassModel` trains three occupancy-detection models. All three assume that the probability of occurrence of a focal species in a particular county is a function of the probability of occurrence in neighboring counties (i.e., a conditional autocorrelation regression model). The three models differ in how the probability of detection given the species is present is determined. The simplest model assumes that the probability of detection given presence is constant across counties. The next model assumes the probability of detection in a county is the same for all counties in a state. This can be justified by examining maps of the density of collections by county—in many cases certain states have very high collection densities, so detectability likely varies by state (e.g., see New Mexico and Kansas in Figure 1). The final model assumes that the maximum probability of detection changes by state, but that the probability of detection in a specific county ranges between a minimum value and this state-wide maximum value.

All three models also estimate a region-wide probability that all detections in a county are mistaken identifications. Misidentification is especially a problem for plants; some estimate that up to a third of all specimens are mis-identified. This value might also be interpreted as the probability that a set of specimens in a county are vagrants or from a disjunct population. Regardless of interpretation, value is “q” in the model output.

The `tropicosMassModel` script trains all three models, then calculates which are statistically sufficient. If at least one model is sufficient, it is retained. If more than one is sufficient, then the model with the lower Watanabe-Akaike Information Criterion (WAIC) is retained as the favored model.