

enmSdmBayes:

A user-friendly occupancy-detection model for modeling herbarium and natural history specimen data at broad biogeographic scales

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

Contents

Introduction	1
What you get	1
More information	2
Downloading and installing	2
Running enmSdmBayes	3
A tutorial: Shapefile input	5
A tutorial: CSV input	11
Troubleshooting	13
Acknowledgements	14

Introduction

Collectively, the world's herbaria and natural history museums hold several billion specimens representing millennia of person-work. This is the basic information on which our knowledge of the distribution of biodiversity is built. Nevertheless, collections are often sporadic in space and time, and as a result, our knowledge of the true distribution of species is strikingly lacking at a time when conservation requires this knowledge. Hence, we need a way to 1) estimate the true distribution of species in an 2) user-friendly manner while 3) correcting for potential spatial and temporal bias in existing collections.

enmSdmBayes fulfills this need. enmSdmBayes is a set of models and a user-friendly interface to these models. Users unacquainted with Bayesian modeling or with running command-line code can use enmSdmBayes to model species' true distributions while correcting for bias in collections. This is a guide for using the interface. Users interested in running the models from the R command-line can browse the scripts in the folder named "R" of this repository. Help on each function is provided in-line with the function.

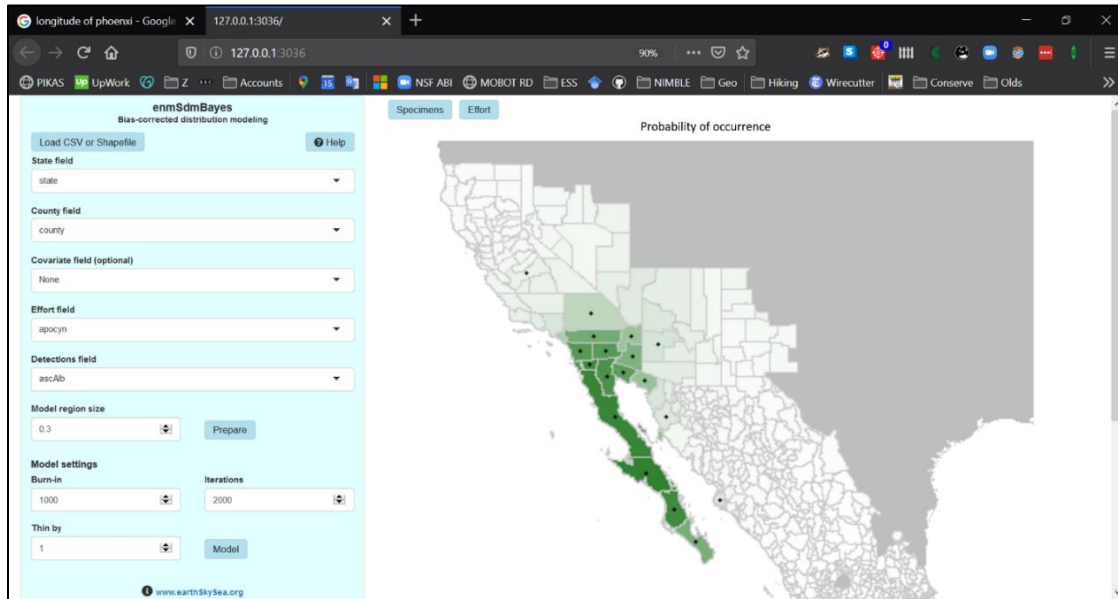
What you get

enmSdmBayes can estimate:

1. The probability of occurrence of a species and uncertainty associated with this probability;

2. The probability of the specie having been collected given that it was present; and
3. The probability that the entire set of specimens in a locale are mistakenly mis-identified as the focal species.

Except for the last, each of these values can be mapped and displayed in enmSdmBayes and exported into shapefile format for processing by GIS software. The interface is simple, easy to use, and runs in your web browser.



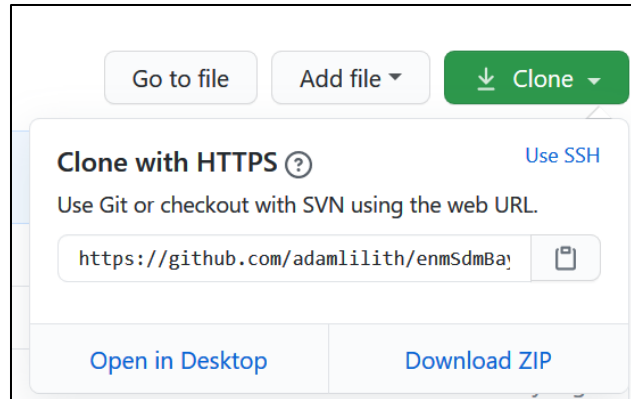
More information

You can get more information on enmSdmBayes from the project website at www.earthSkySea.org/enmSdmBayes and from the project repository <https://github.com/adamlilith/enmSdmBayes>. The latest version of the software will always be available at this second URL.

Downloading and installing

Running enmSdmBayes is fairly easy, but you will need to install some software before you can use it. You may need to have administrative permissions on your computer to use it. The software will run on a Windows, Mac, or Linux machine. All of this software is open-source and free to use.

1. Download and install R (<https://cran.r-project.org/>).
2. Download and install RTools (<https://cran.r-project.org/bin/windows/Rtools/>). You may need to follow directions on this page to get the installation to work correctly.
3. Download and install RStudio (<https://rstudio.com/>).
4. Download the enmSdmBayes repository. The latest version can be found at <https://github.com/adamlilith/enmSdmBayes>. To download the repository, click the “Clone” button then “Download Zip”:



Save the file to your computer and unzip it into a directory.

5. Navigate to the “install” folder within this directory and find the file named “install.r.” Open this in a text editor (or RStudio). Start RStudio and paste the contents of “install.r” into the console pane then hit Enter:

```

> ### install file for enmSdmBayes
### Adam B. Smith | Missouri Botanical Garden | adam.smith@mobot.org | 2020
###
### last updated 2020-03-31

# load/install packages
packs <- c('BIEN', 'coda', 'dismo', 'grpreg', 'grpregoverlap', 'nimble', 'spdep', 'wiqid', 'stringi', 'sp', 'tidyverse',
'raster', 'rgeos', 'scales', 'shiny', 'shinyjs')

for (pack in packs) {
  worked <- do.call(require, args=list(package=pack))
  if (!worked) {
    install.packages(pack, repos='https://cloud.r-project.org')
    do.call(require, args=list(package=pack))
  }
}

pack <- c('omnibus', 'legendary', 'birdsEye', 'enmSdm', 'statisfactory')

for (pack in packs) {
  worked <- do.call(require, args=list(package=pack))
  if (!worked) {
    remotes::install_github(paste0('adamlilith/', pack))
    do.call(require, args=list(package=pack))
  }
}

```

You should not need to do any of these steps again on this computer.

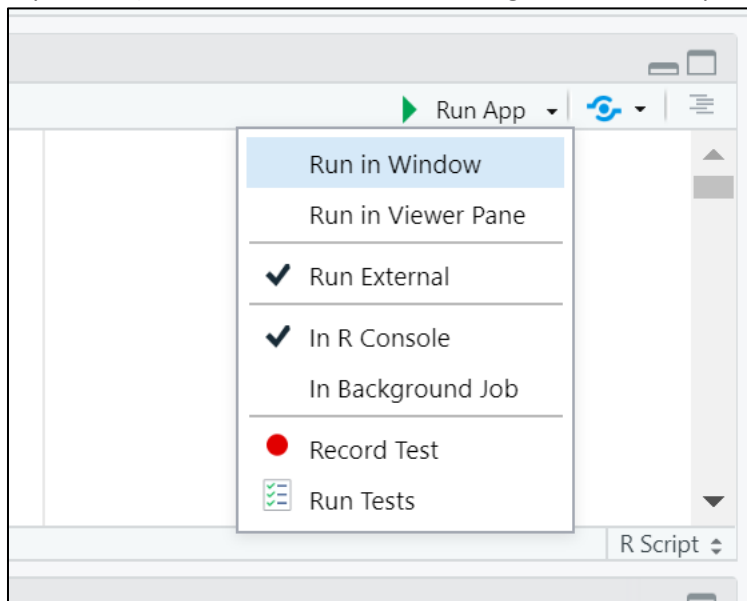
Running enmSdmBayes

1. You are now ready to run enmSdmBayes. In RStudio open the file named “enmSdmBayes.r” in the folder named “R”:

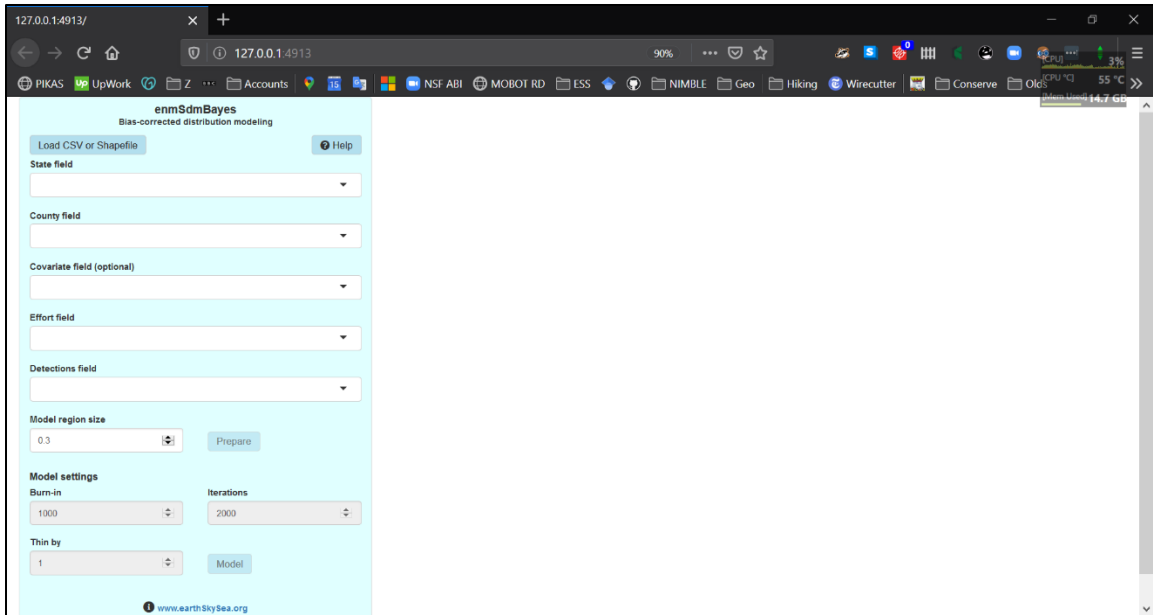
py > Drive > R > enmSdmBayes > R >

Name	Date modified	Type	Size
data	2020-06-28 8:52 PM	File folder	
bayesODMStats	2020-06-18 4:49 PM	R File	2 KB
countryISO3	2020-06-17 11:24 AM	R File	1 KB
darwinCoreSpatialAssign	2020-05-06 10:41 AM	R File	86 KB
enmSdmBayes	2020-06-28 9:58 PM	R File	24 KB
getGeogFocus	2020-06-17 9:42 PM	R File	6 KB
getISO3	2020-06-17 10:39 PM	R File	3 KB
mapBayesODM	2020-06-24 11:30 PM	R File	9 KB
matchDfToGadm	2020-06-17 11:12 PM	R File	3 KB
rhatStats	2020-06-14 12:12 AM	R File	2 KB
trainBayesODM	2020-06-25 4:51 PM	R File	14 KB

We suggest you click the small downward arrow next to the “Run App” button and select “Run External” so that the software runs in your web browser (which in our opinion provides a better experience). You will not need to do this again on this computer unless the setting gets changed.



- Now, click the “Run App” button again. Your web browser should open with the enmSdmBayes app:



Note that if at any time you wish to stop the app you can close the browser window. However, RStudio will continue to run the app in RStudio, so to stop it you can click the “stop” sign:



A tutorial: Shapefile input

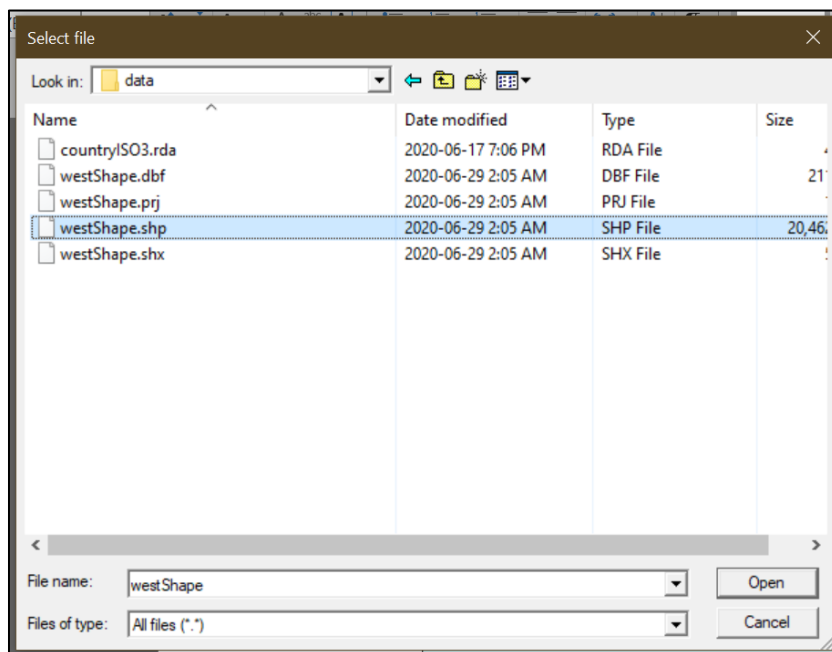
To run enmSdmBayes you need data on collections by county of a focal species plus other species. This data can be represented in either a CSV file, with one row per record, or in a shapefile. The shapefile method is easier, so we will start with that first. The CSV method is described in step 4.

The shapefile must have at least two columns, one reflecting the total number of collections (or “effort”) in each county and one the total number of collections of the focal species (“detections”). For any given row, the number of “efforts” must be \geq the number of detections. You can also have another column with numeric values in it to serve as a predictor (e.g., county area). You can have other columns in the shapefile, too.

The “effort” column should reflect specimens collected in the same region as the focal species and should be numerous. We have found that using specimens from the same family provides a good set of “other” specimens so long as there are sufficient specimens and good geographic coverage. Not all counties must contain specimens.

Collections of the focal species must also have sufficient geographic coverage and number of records. The exact number depends on the situation, but we have found that having at least 100 specimens of the focal species is necessary to produce a sufficient model.

1. To illustrate how to use a shapefile, we will use the “westShape” shapefile in the “data” folder. First, click the “Load CSV or Shapefile” button, then navigate to the “data” folder and select the “.shp” file:



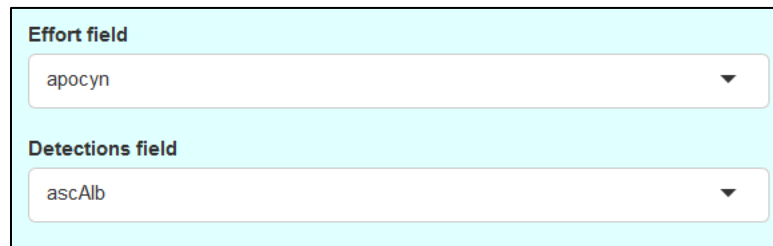
2. Now, select the fields in the shapefile representing “state/province” and “county”:

The form has a light blue background and contains the following fields:

- Load CSV or Shapefile** button (top left) and **Help** button (top right).
- State field**: A dropdown menu with 'state' selected.
- County field**: A dropdown menu with 'county' selected.
- Covariate field (optional)**: A dropdown menu with 'None' selected.

3. We will not use a covariate in this example, but you could choose “areaKm2” from the drop-down menu. The covariate represents a numeric factor you suspect may affect the occurrence of the species.
4. Next, choose the columns in the shapefile that represent “effort” and “detections”. In this example, the column “apocyn” is a tally of the number of collections of Apocynaceae (the dogbane flowering plant family), which we will use as an estimate of collection effort. There are also columns for

number of records of *Asclepias albicans* (field “ascAlb”), *Asclepias californica* (“ascCal”), and *Apocynum cannabinum* (“apoCan”). We’ll use *Asclepias albicans* for this example:



The screenshot shows a light blue rectangular box containing two dropdown menus. The top menu is labeled "Effort field" and has "apocyn" selected. The bottom menu is labeled "Detections field" and has "ascAlb" selected.

5. Now we will select a subset of the shapefile to represent the modeling region. The size of the area is determined by ensuring that all counties with at least one collection of the focal species is included, plus some region around these counties. The size of this region is determined by the “model region size” factor. The default is 0.3, which we have found to work well in a variety of cases. The values should be 0 or greater... to include the entire shapefile use a very large number (e.g., 10 or 20... too large may cause an error, though). Enter a value in the box and click “Prepare”. Depending on the size of the shapefile, it may take a few minutes to complete.



The screenshot shows a light blue rectangular box. On the left, there is a text input field labeled "Model region size" containing the value "0.3". To the right of the input field is a small up/down arrow icon. Further to the right is a blue button labeled "Prepare".

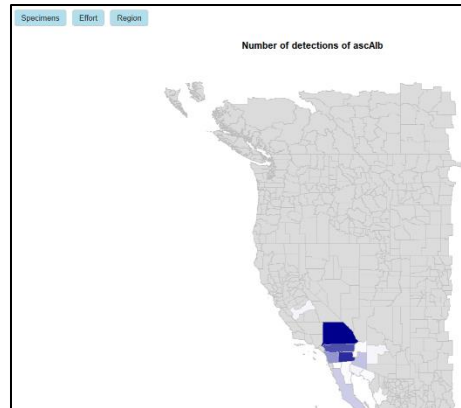
6. When the enmSdmBayes is done generating the modeling region three buttons will appear at the top of the screen:



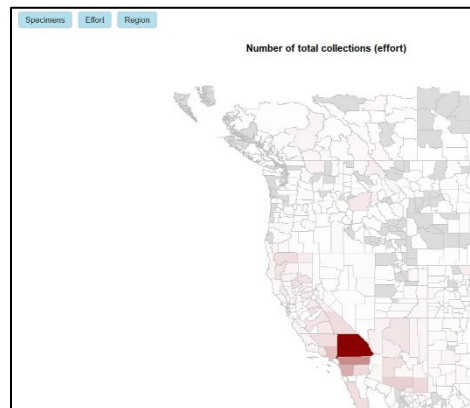
The screenshot shows three light blue buttons with black text, arranged horizontally within a thin black border. The buttons are labeled "Specimens", "Effort", and "Region" from left to right.

Clicking any of these will show a map:

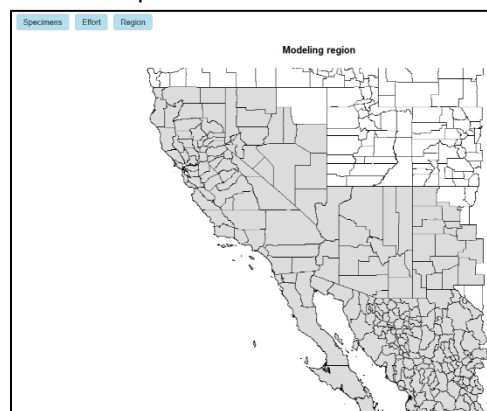
- “Specimens” displays a color-coded map of number of specimens of the focal species in each county. Darker shades indicate more records. Gray indicates no records.



- “Effort” will show a map of the total number of “effort” collections in each county. Darker shades indicate model records. Gray indicates no records.



- “Region” will show the extent of the modeling region (in gray) in relation to the overall shapefile.



7. Finally, we are ready to model the species. The settings for the model are near the bottom of the panel on the left:

Model settings

Burn-in: 1000

Iterations: 2000

Thin by: 1

Model

The underlying model is Bayesian and “solved” using Monte Carlo methods. If you don’t understand this, don’t worry—in effect, it means the model finds the right solution by trying a lot of values while gravitating toward the ones that best fit the data.

Burn-in: The number of iterations the model uses to “teach” itself how to make optimal guesses. Typically, this is at least 1000, but you might try larger numbers (2000, 5000, 10000) if the model doesn’t seem to be working well.

Iterations: Total number of interactions the model tries, including burn-in. This number must be greater than the value for burn-in. Typically 1000 iterations *after burn-in iterations* is the smallest viable number, but you may want to try larger numbers (2000, 5000, 10000, ... , even maybe a million or more).

Thin-by: To save memory, you can retain every “thin-by” iteration and discard the rest. This is helpful if you have to run the model for many iterations and your computer’s memory becomes full. It provides no other benefit. The total number of samples is (iterations – burnin) / thin. Again, you want at least 1000 samples for a good model, so select values with this in mind.

A note on run-time: `enmSdmBayes` is not like other species distribution models you may have heard about (e.g., Maxent). It corrects for bias and estimates the probability of occurrence using only collections data, plus (optionally) a covariate. In contrast, other distribution models do not correct for bias and require several, sometimes many predictors. This requires some heavy thinking about what the predictors should be and a lot of extra work to acquire and prepare the predictors. And since we have so little definitive knowledge about what really shapes any particular species’ range, it is often advisable to try several combinations of predictors. In sum, this means that traditional species distribution modeling takes a lot of work to do it well!

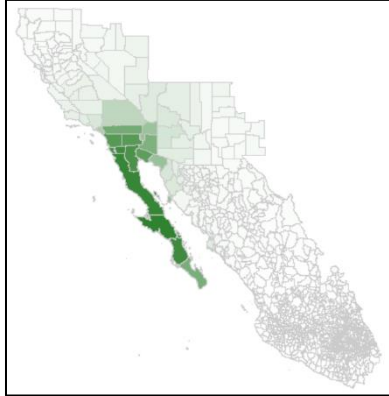
`enmSdmBayes` is much simpler in the time required of you. However, as a result, it can take the *computer* a lot of time to create a viable model—depending on the amount of data and settings for `niter`, `nburnin`, and `thin`, sometimes hours or even days. Even if you set `niter` and `nburnin` to unrealistically small values, the computer will take a few minutes to compile the functions. So sit back, take a break, and be thankful it’s using the computer’s time—not yours.

When you are ready, click the “Model” button!

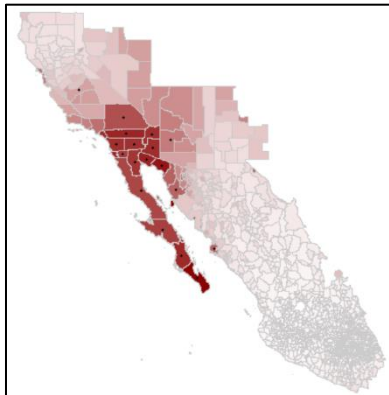
8. When the model is done some extra buttons appear at the top.

“Occupancy”: Probability of occurrence (psi). Darker shades correspond to a higher probability of occurrence. You will notice that the probability of occurrence may be low

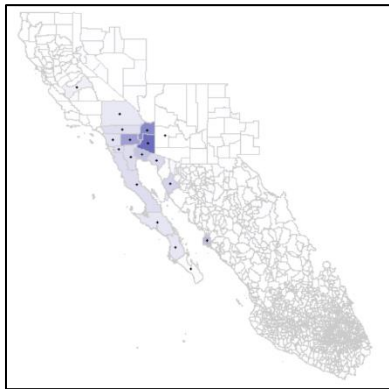
in a county in which we have specimens of the species. This happens in most models—not every occurrence is predicted equally well. *enmSdmBayes* also tends to predict occurrence probability will be higher closer to known occurrences. So occurrences that are far from others may have a low probability of occurring (which makes sense).



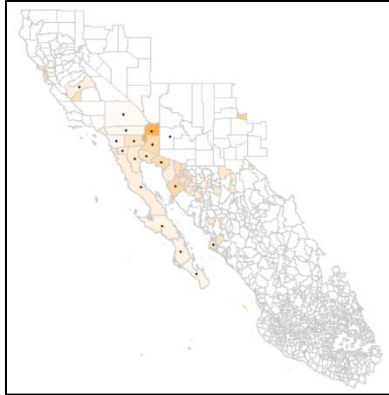
“Occ. uncert.”: 90% credibility intervals around the probability of occurrence (i.e., uncertainty in occurrence probability. Called “psi90CI” in the shapefile (see below).



“Detection”: Probability of having collected the species *assuming* it is present in that county given the observed collection effort (p).



“Detect. uncert.” 90% credibility intervals around the probability of detection. Called “p90CI” in the shapefile.



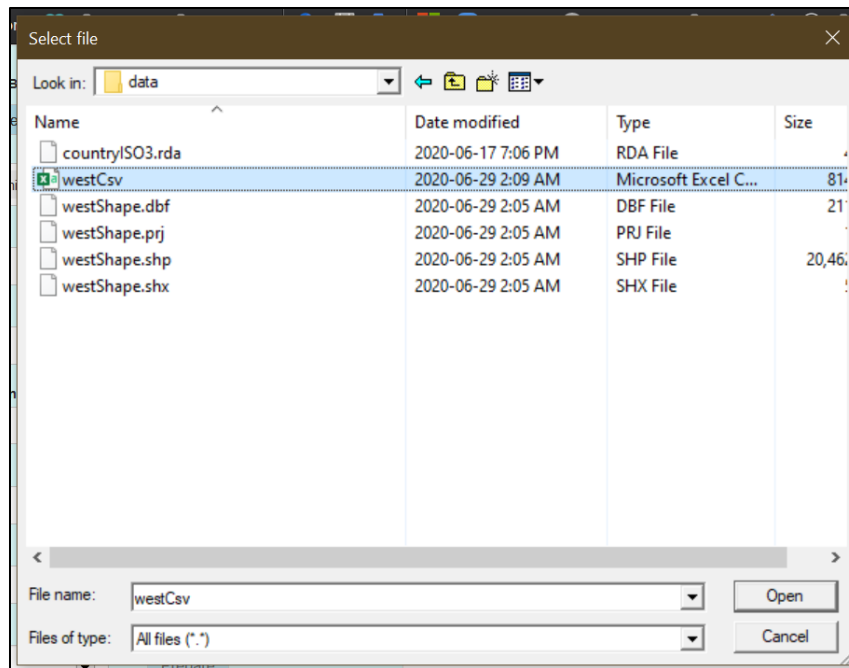
“Statistics”: Statistics relevant to the modeling situation.

9. When you are satisfied with your model, you can save the output as a shapefile. The file is saved automatically in the “data” folder. Note that if you create another model for the same or a different species this file will be overwritten, so it is wise to move it out of this folder if you want to keep it!

A tutorial: CSV input

CSV method: The CSV should have one row per record with the following columns:

- Country names. enmSdmBayes will try to match the country names to spatial data on the web, but to do so the names must be interpretable. enmSdmBayes will try common variants (e.g., “United States”, “USA”, or “Estados Unidos”), but cannot match every variant, so if the following step does not work, you may need to clean the country names.
 - State/province where specimen was collected. Names will be matched as well as they can, but using standard names (i.e., no abbreviations) will work best.
 - County where the specimen was collected. Again, these should be as “standard” as possible.
 - Species name. You will need collections from more than one species to model any particular species. The more “other” species there are, the better. The “other” specimens are used to estimate the amount of collection effort expended in a county. You can remove duplicates beforehand if you believe duplicates mis-estimate actual effort (though collection of duplicates does require some effort).
1. To illustrate the CSV method we will use the “westCSV.csv” file in the “data” folder. Click the “Load CSV or Shapefile” button and load this file:



- Next select the name of the column with the names of countries:

Load CSV or Shapefile
Help

Country field

In this case, a warning message will appear to let you know that a country that could not be identified appears in that column. The countries that appear in that column (plus any invalid ones) are shown:

Country

Canada Mexico United States NA?

You can delete the ones you do not want to include in the modeling region (and you should delete any invalid ones).

Country

Canada Mexico United States NA?

Clicking anywhere in the white part of the box will show you a list of valid country names. If you have specimens that appear in a country but that country is not selected in the box, you may have to clean the CSV manually so the country name(s) match the accepted names in the box.

- Select the columns representing state/province and county. Leave the covariate field blank.

Country

Canada Mexico United States

State field

stateProv

County field

county

Covariate field (optional)

None

- Now, we'll tell enmSdmBayes the name of the column with species names in it. This column will be used to tally effort (total number of collections in a county) and number of detections of the focal species, so we'll select the same column for the "Effort" and "Detections field" boxes. In our case this column is named "binomial".

Effort field

binomial

Detections field

None

When we select the "Detections field" column the name of the field is replaced by a list of all the species in that column.

We will select *Asclepias albicans* as our species.

Detections

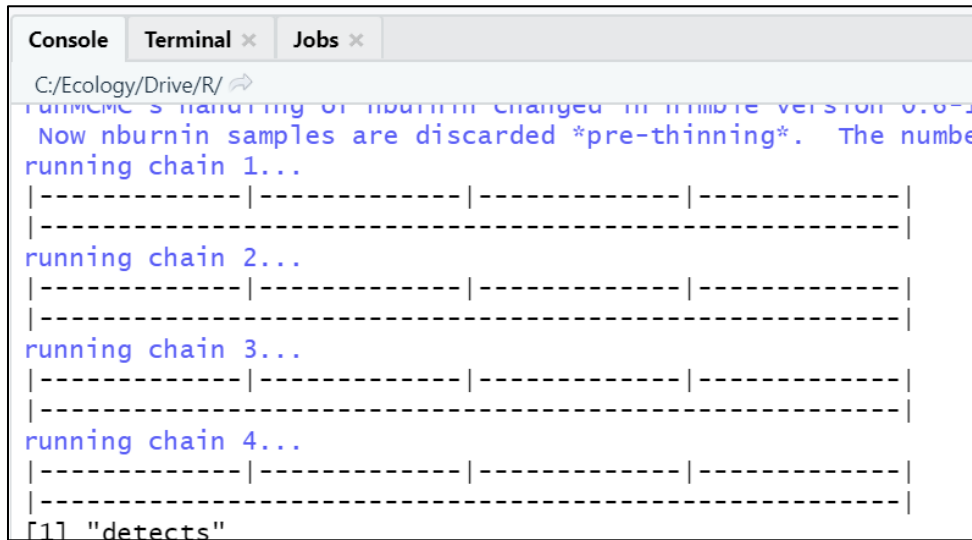
Asclepias albicans

- Now, click the "Prepare" button. For more information on the "Model region size", see step 5 in the previous section on using a shapefile as input. enmSdmBayes does three things when you click this button and you have loaded a CSV file. First, it downloads shapefiles for the selected countries. Second, it attempts to match the records to states/provinces and counties in the shapefile to create a record of "effort" and specimens of the focal species. Finally, it creates a modeling region.
- At this point the CSV method is the same as the method using shapefiles, starting at step 6 in the previous section on using a shapefile as input, so please go there and continue as described.

Troubleshooting

- Nothing seems to be happening: You can see messages from the program by scrolling down to the bottom of the page.

You can also see more details about what is happening when you click a button by viewing the messages in RStudio. These messages are not essential, but they tell you if something is being processed.



```
Console Terminal x Jobs x
C:/Ecology/Drive/R/
runMCMC's handling of nburnin changed in rjags version 0.8-2
Now nburnin samples are discarded *pre-thinning*. The number
running chain 1...
|-----|-----|-----|-----|
|-----|-----|-----|-----|
running chain 2...
|-----|-----|-----|-----|
|-----|-----|-----|-----|
running chain 3...
|-----|-----|-----|-----|
|-----|-----|-----|-----|
running chain 4...
|-----|-----|-----|-----|
|-----|-----|-----|-----|
[1] "detects"
```

Acknowledgements

This software and the modeling algorithms on which it was built were originally formulated by Dr. Camilo Sanín. We gratefully acknowledge feedback from Drs. Stephen Beissinger and Perry de Valpine of UC Berkeley and assistance with the shiny app from Chris Gerdji of Helgasoft. This project was made possible in part by the Institute of Museum and Library Services National Leadership grant to ABS (FAIN MG-30-15-0094-15).