Day 3: Introduction to Ecological Niche Modelling with *MaxEnt* and *Wallace*

Eugenio Valderrama & Jesus Martínez

26 Oct 2020

Abstract In this session we will explore some methods for potential distribution models or ecological niche models. Several repositories are now available to obtain occurrences for many organisms as well as environmental data with varying degrees of spatial resolution. We can combine those occurrences and environmental variables to create models that allow us to predict the potential distribution of species. We will use *MaxEnt* algorithm and environmental variables related to temperature and precipitation from *Worldclim* (by far the most popular algorithm and variables for these analyses). A very important consideration for these analyses, is to mind the quality and format of the data. We will use *R* package *Wallace* to obtain and curate occurrences and the climatic variables in an easy and interactive way instead of having to install and use more intricate GIS (geographic information systems) applications that are demanding in terms of computer requirements and sometimes have expensive licenses (e.g. *ArcGIS*). *Wallace* package is very comprehensive and we will be just scratching the surface of its possibilities. Because *MaxEnt* is an independent java application getting *R* to run it is different on every machine (and normally requires extensive troubleshooting), we decided to use just the first steps of the Wallace pipeline. However, if you are interested check their website and get the most of the GUI (graphical user interface), that even allows you to get all the *R* code you used in your analyses to process your data more efficiently and in a replicable fashion or in a server without a GUI (like you did on the last sessions!... we also have a full tutorial that only uses *R* if you are interested).

Getting occurrences and climatic variables with *Wallace*

First open *Rstudio* (or *R* if you are an old school rebel) and install and load *Wallace* package with the following commands.

```
install.packages('wallace')
library('wallace')
run_wallace()
```

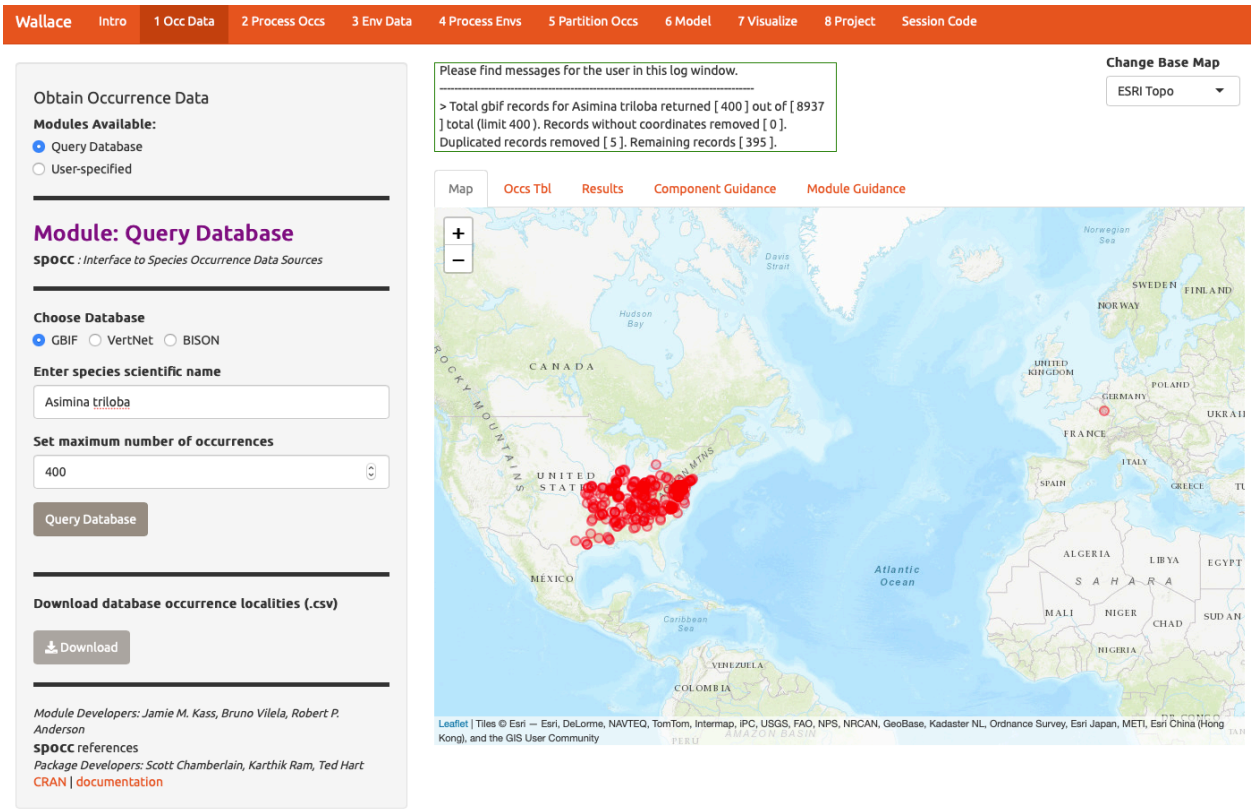
A *shiny* app (a package to program applets on R) will start on your internet browser, this means *R* is controlling your browser to create a GUI. If you close *R* or your browser, you will terminate the session and will have to start over with the run_wallace() command. You can refresh your browser to restart the applet (you will lose all the info!).

Once in *Wallace* applet you can check the tabs in the banner on top. We will only use tabs 1. to 4.

1. **Occ Data** tab

The structure of the tabs will be similar, we will mainly use the **Map** display. If you want to get documentation and help go to the **Model Guidance** display that will be different in each tab and module. In the **Occ Data** tab we will download occurrences from the *GBIF* repository (that compiled data from natural history collections across the globe and initiatives like *iNaturalist*).

Click the **Query Database** button and choose the **GBIF** database. Enter the scientific name of the organism you want to query for, for this example use Asimina triloba, the Pawpaw tree. This tree is native of the eastern United States and Canada and a very interesting species of the mainly tropical family Annonaceae. Pawpaw trees have super tasty fruits that were eaten by Native Americans, look like a tropical rainforest bush, have super interesting pollination biology and grow on campus! (ask Eugenio if you want to try finding them). Set the number of occurrences to 400. Hit the **Query Database** button and you should get the records on the map. You can click on the points to check the details. Do you see occurrences that don't make any sense?



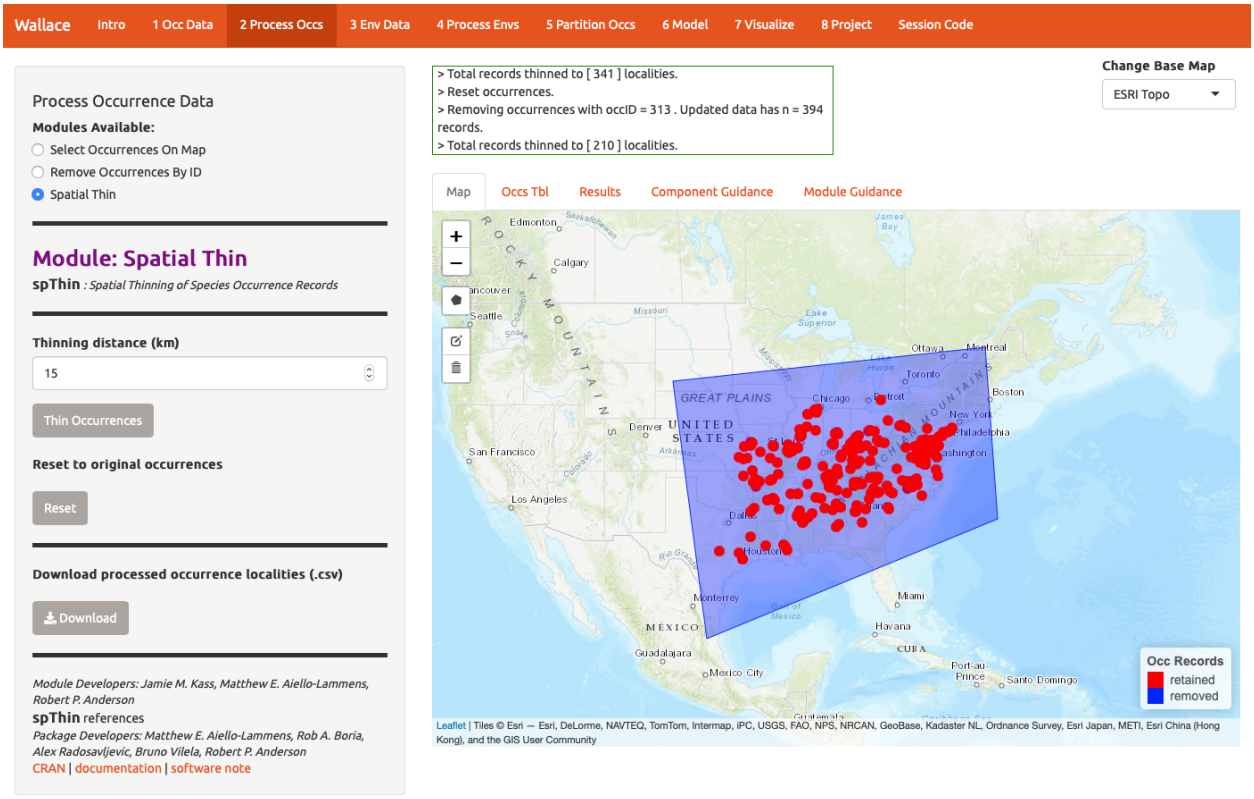
2. **Process Occs** tab

If you obtained points that don't make any sense (a tree in the middle of the ocean or in the wrong continent) they could be problematic downstream with the algorithms or even worse, leak through the workflow and make your model unreliable. You can filter the appropriate references on the map by selecting what to keep over the map with a polygon (**Select Occurrences On Map**) or deleting by ID (**Remove occurrences by ID**; you can check each ID by clicking on the records on the map). Filter your data using the ideas of the lecture. Don't panic if you messed it up, just use the **Reset** button to start over. Be sure to use the **Select Occurrences** button if using a polygon or **Remove Occurrence** when using the ID's.

After selecting your occurrences apply **Spatial Thin**. This step allows us to reduce some of the biases in sampling and also to have a better idea of the amount of data for the resolution of the climatic data we are going to use. Use a **Thinning distance** of 15 km and check if you still have

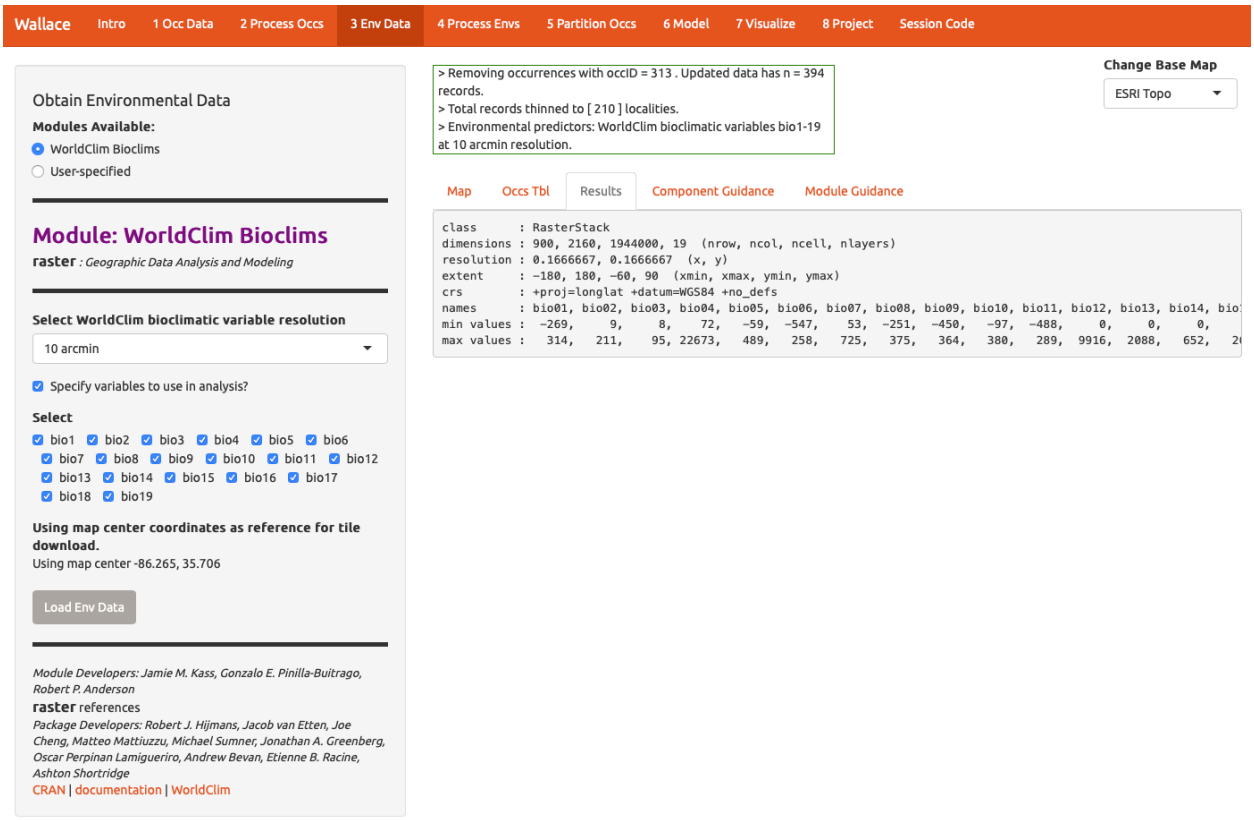
some records (check the text box below the panel with the tabs). We will use a course resolution for the climatic variables to run analyses quickly (10 arcmin at a latitude of 40 arcdegrees or *c.*14km^2 grid for the climatic variables), but have in mind you should adjust this step to the resolution of the climatic data you are going to use and to the biology of the organism of interest. When ready click on **Thin Occurrences.**

Once you are ready, download your processed and thinned occurrences as a .csv file and save it in a new folder named 'myENM' in your local drive. You can check that file in the text editor you installed for last sessions (*Sublime Text* or *BBEdit*). You should see in the first column of that text file separate by commas the name of the species (should be the same for all records) and the geographic coordinates (longitude and latitude in decimal degrees format) as well as info for the record. Do you think *iNaturalist* data is reliable? Would be better to use just herbarium specimens?



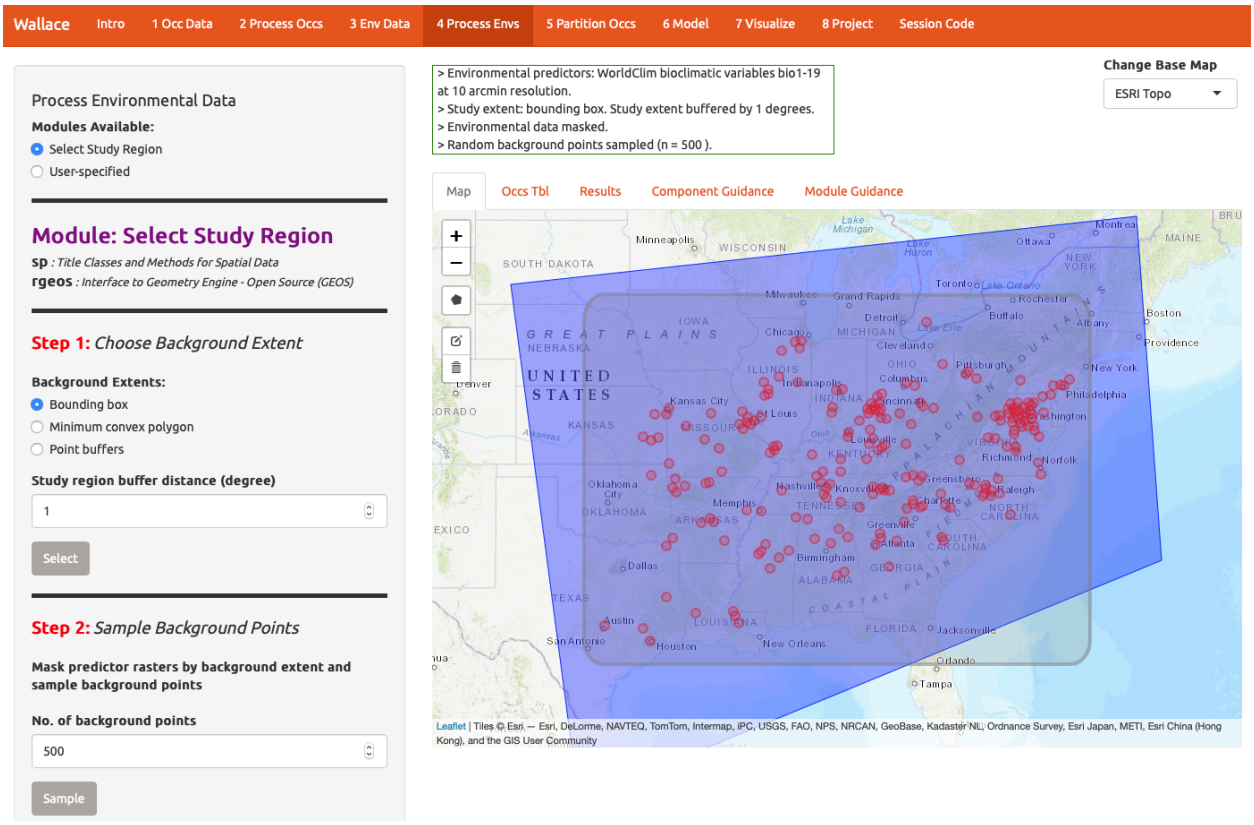
3. **Env Data** tab

In this tab we will get the data of the environmental variables. Check the **WorldClim Bioclims** variables button and select the **10 arcmin** resolution. You can select which ones to use with the **Specify variables to use in the analysis?** box. We will use the whole 19 variables but there is debate about if you should remove correlated variables in your study area and which method to decide which should be removed. Click on **Load Env Data**.

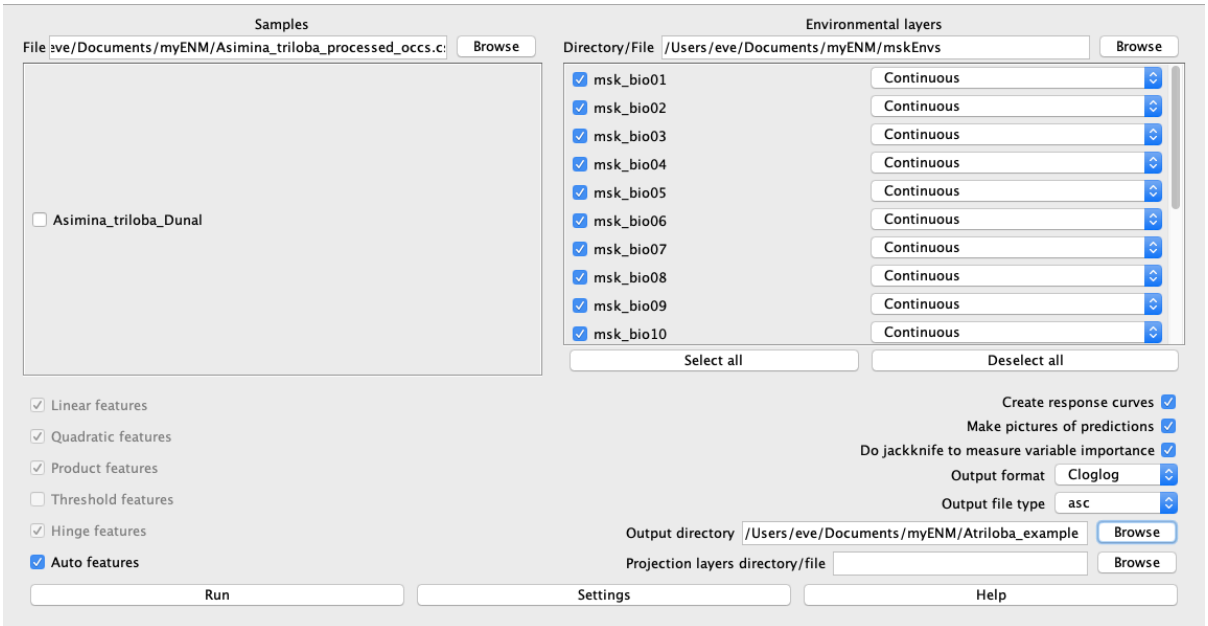


4. **Process Envs** tab

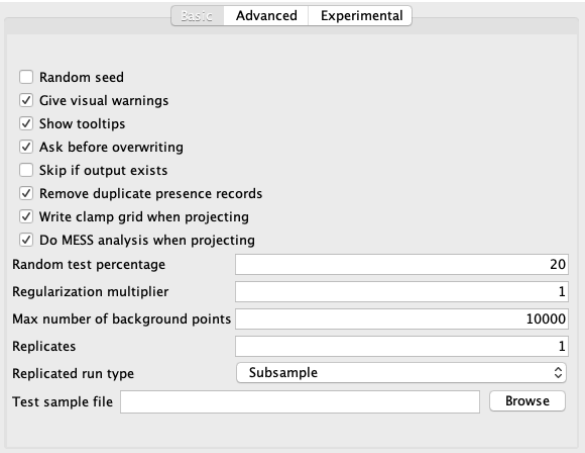
Focusing on a specific study area is very important because of the pseudoabscences being used in the analyses. We assume that the area we include was evenly sampled and that the climate is the main factor preventing the species to inhabit the areas where occurrences don't exist. We will use the simplest approach that is selecting as study region a bounding box around our occurrences with a buffer distance of 1 arcdegree and keep the climatic variables for that area only. Use the **Select** button after clicking the options **Select Study Region**, **Bounding box** and setting the **Study region buffer distance (degree)** to 1. We are not using the **Step 2** because we will just download our climatic data, but we need to do that in order to be able to export in the app, so click on **Sample** after specifying 500 as **No. of background points**. You just prepared the environmental variables, now **download** the climatic data in **ASCII** format to the myENM folder you created. Navigate to the myENM folder and unzip the file with the data you just downloaded. If you have a .csv file with your georeferenced occurrences and a folder with the *bio* variables in ASCII format, we are done with *Wallace* and you can close the browser window and *R*.



Now go to the directory where you downloaded *MaxEnt* and execute maxent.java. Select your .csv file on the **Samples** section and the whole directory with the prepared climatic variables in the **Environmental layers** section. Check the **Create response curves**, **Make pictures of predictions** and **Do jackknife to measure variable importance** boxes. Create a new directory within myENM called 'Atriloba_example' and specify it as the **Output directory**.



Go to the **Settings** window and set **Random test percentage** to 20 and **Replicated run type** to **Subsample** on the **Basic** tab. These changes set the analysis to exclude a random subset of 20% of your occurrences of the fitting of the model and use those independent observations to test how good the model is. As you can imagine the latest settings imply you are losing some of the starting occurrences, but it is the more conservative way to assess the model's performance (there are other alternatives when you have too few occurrences). Close the **Settings** window and **Run**. You will get some warnings regarding some unused columns on your csv file and *MaxEnt* will hold until you close them (if you want to avoid these messages just leave the first three columns 'name', 'longitude' and 'latitude' on your csv file… I think it is quicker just to disregard those warnings). The algorithm should be quick and will generate a .html file on your output directory with all your results and accompanying text that we checked through at the end of the lecture.



Homework:

1. You should provide the resulting *html* file for each model and a document with the answers to the following questions (read it all through before starting):

a. Pick an organism of interest (and pick one with a good number of occurrences) and create a model in *MaxEnt* with climatic variables (at least 200 occurrences after filtering) and answer the following questions. How did you filter the occurrences and why?

b. How did you select the study area and why?

c. Is your model reliable when considering the ROC curve?

d. Which variables were more informative for your analysis, does it make sense according to the biology of the organism you picked?

e. What are the potential flaws of your resulting model? (think in the lecture and in your organisms!)

2. Create a model for the same organism of the previous question but this time use only the occurrences from natural history collections. Your GBIF occurrences have a column called 'basisOfRecord', if it says 'PRESERVED_SPECIMEN' that observation is linked to a museum specimen. Filter the data of your csv file (in a text editor or *excel*) being careful not to change the format of the file. You may have to increase the maximum number of observations obtained in *Wallace* to match the 200 of the first attempt (try to include a similar number of occurrences). Compare the two models and discuss your results using the questions you answered for the first one.