Using Digitized Collections-Based Data in Research: Applications for Ecology, Phylogenetics, and Biogeography

Sponsored by iDigBio and BiotaPhy

Florida Museum of Natural History, University of Florida
and University of Kansas

The following are hands-on exercises to introduce the participants to the programs and protocols described during the workshop.

Table of Contents

SCHEDULE

Times are in EDT

10:00	Welcome and Overview of the Workshop – Pam
10:15	iDigBio Portal – Pam
10:30	Data Downloads – Lauren
10:50	Break
11:05	Data Cleaning – Shelly
11:35	Georeferencing – Andre
12:00	Break
12:15	Climate Processing – Andre
12:30	Climatic Niche – Shelly
1:00	Lunch
1:30	***Question Session***
2:00	Applications of ENMs – Pam
2:15	Ecological Niche Models – Maria
2:30	Interpreting ENM Results – Maria
2:50	Post-ENM analysis – Shelly
3:10	***Question Session***
3:30	Break
3:50	Phylogenetic Diversity – Doug/Maria
4:15	Intro to BiotaPhy – Doug
4:30	BiotaPhy – Maria
4:45	***Question Session***
5:00	End

Workshop Leaders:
Pam Soltis: psoltis@flmnh.ufl.edu
Doug Soltis: dsoltis@ufl.edu
Maria Cortez: mariacortez@ufl.edu
Shelly Gaynor: michellegaynor@ufl.edu
Andre Naranjo: aanaranjo@ufl.edu
Lauren Whitehurst: laurenwhitehurst@ufl.edu

SET-UP

(1) Download the dropbox file locally (suggested "Desktop/")

- https://www.dropbox.com/sh/6sxqnqodxv58mp4/AAAPjftr1UDVu3wK7jC0kaXJa?dl=0

(2) R and R Studio (demo built using R Versions 3.5 and 3.6)

https://www.rstudio.com/products/rstudio/download/
https://cran.rstudio.com/index.html

- Download and install R and the free desktop version of RStudio
    - Then in the shared dropbox folder (or the version of this folder that you downloaded):
        - Open the R project by double clicking the .Rproj file. This can be found under "Demo/Rbased/CrashCourse/CrashCourse.Rproj"
            - Navigate to 00_Set-up.R. Click on 00_Set-up.R; then, to install the files that you will need, go to Source in the upper left quadrant and select Source with Echo from the drop-down menu. The packages will be installed automatically.

(3) QGIS

- QGIS (version: 2.18.23 – MUST BE v. 2.18.23)
    - macOS: https://qgis.org/downloads/macOS/QGIS-OSX-2.18.23-1.dmg
    - windows:
        - https://qgis.org/downloads/QGIS-OSGeo4W-2.18.23-1-Setup-x86.exe
            - https://qgis.org/downloads/QGIS-OSGeo4W-2.18.23-1-Setup-x86.exe.md5sum
        - 64 bit - https://qgis.org/downloads/QGIS-OSGeo4W-2.18.23-1-Setup-x86_64.exe
            - https://qgis.org/downloads/QGIS-OSGeo4W-2.18.23-1-Setup-x86_64.exe.md5sum

(4) BiotaPhy

- Exercise: The **data for an additional exercise** can be found at "/Demo/Biotaphy/BiotaPhy_Hands_On_Sub.zip"
    - The **results** can be found at (2) "/Demo/Biotaphy/Results_Demo.zip"
        - Please rename files using the following format: LASTNAME_FIRSTNAME.extension (csv for occurrence data, tre for the phylogeny, zip for hypotheses). The BiotaPhy server requires that jobs and files have unique names. If your file name is already in use on the server, you will receive a notification asking you to change the filename and try again.

DATA DOWNLOAD

(A) iDigBio web-portal (https://www.idigbio.org/portal/search)

- Download data from your web browser

(B) R based

- Open the R project by double clicking the .Rproj file. This can be found under "*Demo/Rbased/CrashCourse/CrashCourse.Rproj*"
  - Navigate to 01_*Download_Occurrence_Data.R*
- Or follow along on the *CrashCourse_2020.html* file which can be found *"Demo/Rbased/CrashCourse_2020.html"*. This file can be opened in a web-browser.

DATA CLEANING

**Depending on the size of your dataset and your comfort with R and RStudio, you may or may not want to use the R script that we provide. The basic steps are as follows:

1. Resolves taxon names
2. Rounds up the latitude/longitude to our desired coarseness and removes points that are not precise enough
3. Removes coordinates at 0.00
4. Removes coordinates in cultivated zones, botanical gardens, etc.
5. Removes duplicates
6. Removes coordinates outside of our desired range
7. Produces a csv file with just the latitude and longitude for each record

(A)Manual – Optional

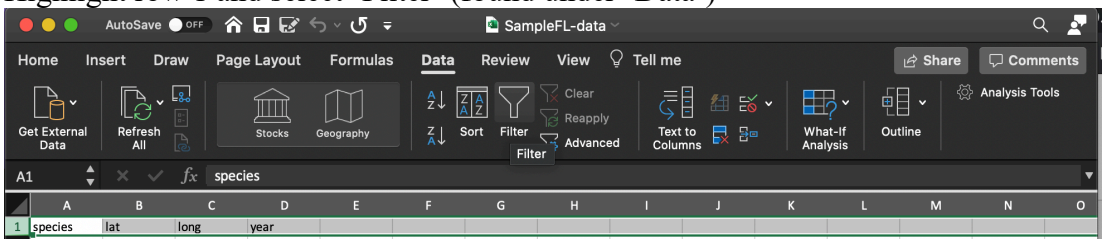
For another excel based demo, checkout: Gaynor, M. (2020). Cleaning Biodiversity Data: A Botanical Example Using Excel or RStudio. Biodiversity Literacy in Undergraduate Education, QUBES Educational Resources. doi:10.25334/DRGD-F069

Using *"Manual/Data_Cleaning/SampleFL-data.csv":*

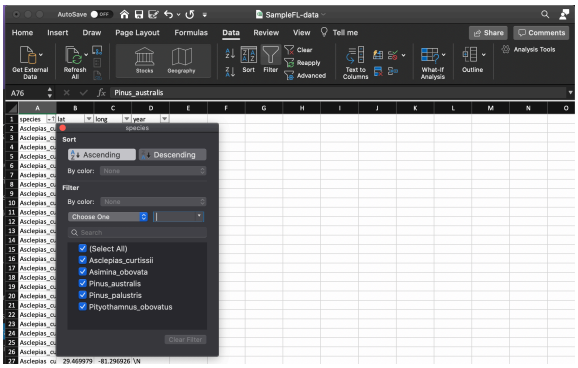
1. Resolve taxon names

1a. Filter and sort columns

Highlight row 1 and select 'Filter' (found under 'Data')



Sort by species



1b. Identify which names are synonyms

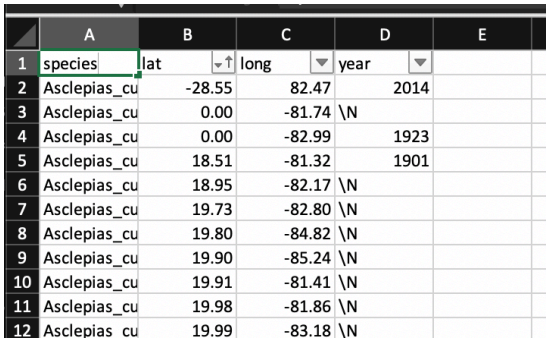
- Many resources to do this:
  - Encyclopedia of Life
  - Catalogue of Life
  - iPlant Taxonomic Name Resolution Service
  - PLANTS database
  - Integrated Taxonomic Information System (ITIS)
  - The Plant List
  - Primary literature

species	new
Asclepias_curtissi	Asclepias_curtissi
Asimina_obovata	Asimina_obovata
Pinus_australis	Pinus_palustris
Pinus_palustris	Pinus_palustris
Pityothamnus_obovatus	Asclepias_curtissi

1c. Create a column titled ‘new’ and replace with the accepted name.

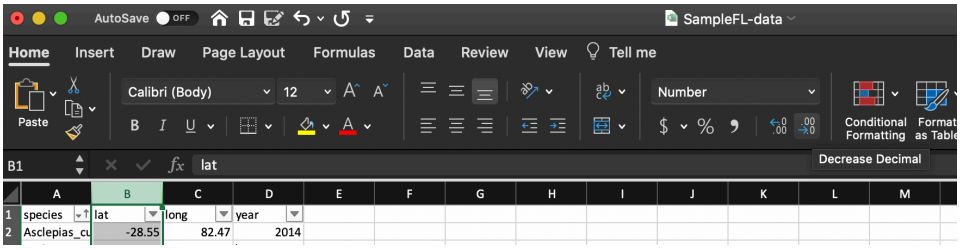
	A	B	C	D	E	F
1	species	lat	long	year	new	
2	Asclepias_cu	-28.55	82.47	2014	Asclepias_curtissii	
3	Asclepias_cu	29.47	0.00	\N	Asclepias_curtissii	
4	Asclepias_cu	28.71	-51.23	2000	Asclepias_curtissii	
5	Asclepias_cu	28.30	-70.70	1955	Asclepias_curtissii	
6	Asclepias_cu	27.08	-80.40	2000	Asclepias_curtissii	
7	Asclepias_cu	26.64	-80.44	1981	Asclepias_curtissii	
8	Asclepias_cu	27.38	-80.44	2003	Asclepias_curtissii	

1d. Replace “species” column with the “new” column and rename to “species”



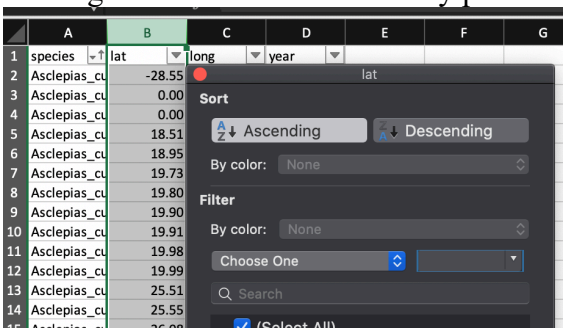
	A	B	C	D	E
1	species	lat	long	year	
2	Asclepias_cu	-28.55	82.47	2014	
3	Asclepias_cu	0.00	-81.74	\N	
4	Asclepias_cu	0.00	-82.99	1923	
5	Asclepias_cu	18.51	-81.32	1901	
6	Asclepias_cu	18.95	-82.17	\N	
7	Asclepias_cu	19.73	-82.80	\N	
8	Asclepias_cu	19.80	-84.82	\N	
9	Asclepias_cu	19.90	-85.24	\N	
10	Asclepias_cu	19.91	-81.41	\N	
11	Asclepias_cu	19.98	-81.86	\N	
12	Asclepias_cu	19.99	-83.18	\N	

2. Rounds up the latitude/longitude to our desired coarseness and removes points that are not precise enough. You can round by using ‘Decrease Decimal’.

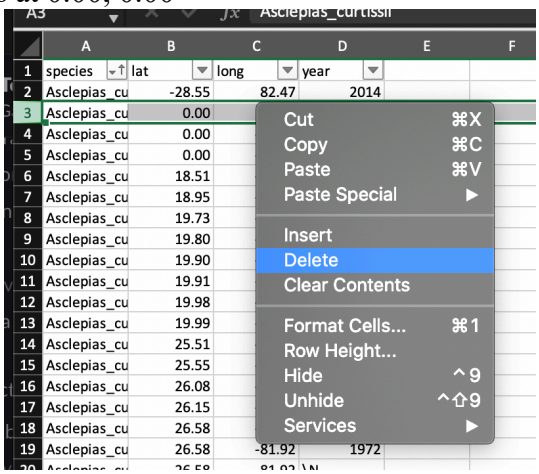


3. Removes coordinates at 0.00

3a. Filter/Sort the ‘long’ or ‘lat’ column to identify points at 0.00, 0.00

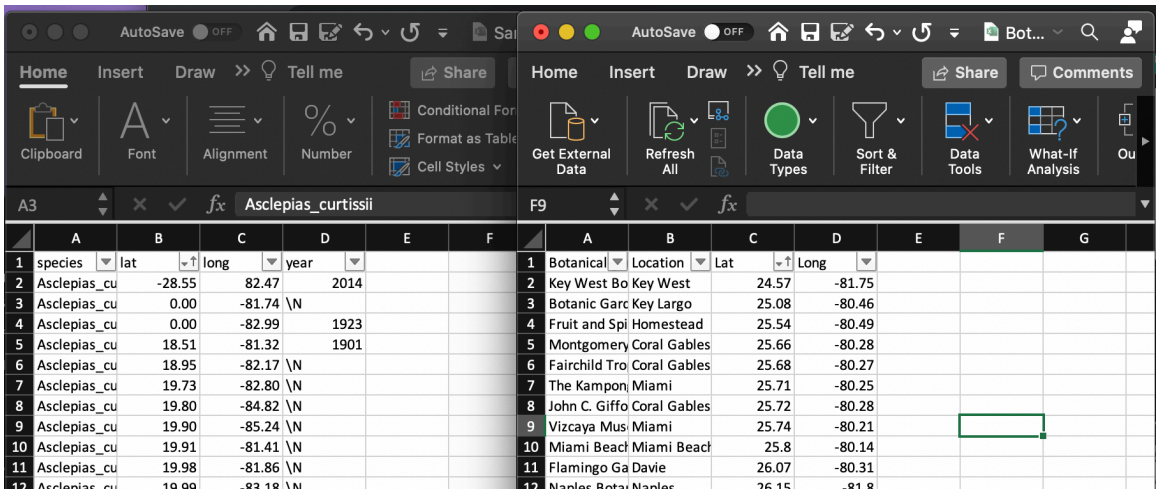


3b. Delete points at 0.00, 0.00



4. Removes coordinates in cultivated zones, botanical gardens, etc.

4a. Open “BotanicalGardensFloridaCoordinates.csv” and Filter/Sort “Lat”. Then manually compare the two files and delete any points in “SampleFL-data.csv” that are shared with “BotanicalGardensFloridaCoordinates.csv”



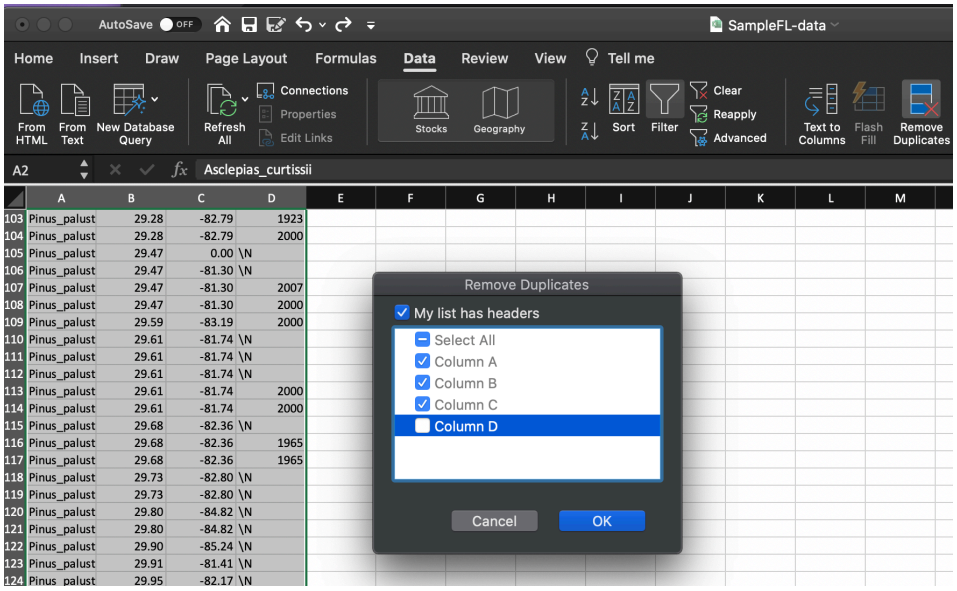
	A	B	C	D
1	species	lat	long	year
2	Asclepias_cu	-28.55	82.47	2014
3	Asclepias_cu	0.00	-81.74	\N
4	Asclepias_cu	0.00	-82.99	1923
5	Asclepias_cu	18.51	-81.32	1901
6	Asclepias_cu	18.95	-82.17	\N
7	Asclepias_cu	19.73	-82.80	\N
8	Asclepias_cu	19.80	-84.82	\N
9	Asclepias_cu	19.90	-85.24	\N
10	Asclepias_cu	19.91	-81.41	\N
11	Asclepias_cu	19.98	-81.86	\N

	A	B	C	D
1	Botanical	Location	Lat	Long
2	Key West Bo	Key West	24.57	-81.75
3	Botanic Garc	Key Largo	25.08	-80.46
4	Fruit and Spi	Homestead	25.54	-80.49
5	Montgomery	Coral Gables	25.66	-80.28
6	Fairchild Tro	Coral Gables	25.68	-80.27
7	The Kampon	Miami	25.71	-80.25
8	John C. Giffo	Coral Gables	25.72	-80.28
9	Vizcaya Mus	Miami	25.74	-80.21
10	Miami Beach	Miami Beach	25.8	-80.14
11	Flamingo Ga	Davie	26.07	-80.31

5. Removes duplicates

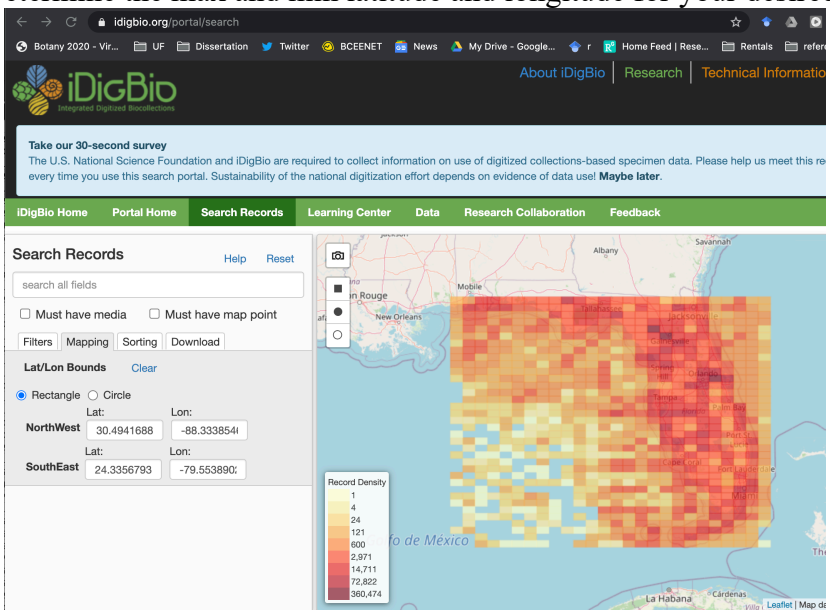
5a. ‘Data’ -> ‘Remove Duplicates’

- Select columns species, lat, and long



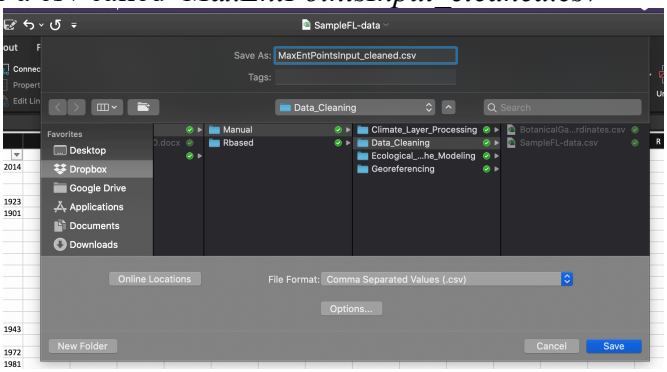
6. Removes coordinates outside of our desired range

6a. Determine the max and min latitude and longitude for your desired range.



6b. Remove points that fall outside that range.

7. Save cleaned file a csv called *'MaxEntPointsInput_cleaned.csv'*



(B) R based

- Open the R project by double clicking the .Rproj file. This can be found under *"Demo/Rbased/CrashCourse/CrashCourse.Rproj"*
    - Navigate to 02_*Occurrence_Data_Cleaning.R*
- Or follow along on the *CrashCourse_2020.html* file which can be found *"Demo/Rbased/CrashCourse_2020.html"*. This file can be opened in a web-browser.

GEOREFERENCING

(A) Manual

Files for this activity can be found in the "Demo/Manual/Georeferencing/" folder.

Georeferencing Demo

Resources:

GeoLocate: http://www.museum.tulane.edu/geolocate/web/default.html
GeoLocate – Web application: http://www.geo-locate.org/web/WebGeoref.aspx
Google Maps: https://www.google.com/maps
Falling Rain: http://www.fallingrain.com
Getty Thesaurus of Geographic Names (TGN): http://bit.ly/Getty-TGN
Fuzzy Gazetteer: http://dma.jrc.it/services/fuzzyg/

1. Use the **standard** GeoLocate client to identify the first three localities in the GeorefExamples_Florida.xls file.
    a. Enter the locality string, country, state, and county information from the Excel sheet.
    b. Click "Georeference."
    c. Inspect the "Possible Locations" by clicking on the "XX possible locations found" where XX is the number of locations GeoLocate identified.
    d. Use an alternative resource to double check the locality. Try Google Maps.
    e. Adjust the point location as you see fit. The green point is the active one.
    f. Click the green point on the map, then click "Edit uncertainty". Adjust the uncertainty radius by moving the grey arrow.
    g. Return to the "Workbench" and record the latitude, longitude, and uncertainty.
        i. If the uncertainty is >1000 then discards the points.
2. Optional.
   Use the **batch** GeoLocate client to upload the localities in the GeorefExamples.xls file.
    a. Copy and paste the appropriate information from the GeorefExamples.xls file into your own GeoLocateBatchFormat.csv.
        i. http://www.geo-locate.org/standalone/tutorial.html

	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10
Row1	locality string	country	state	county	latitude	longitude	correction status	precision	error polygon	multiple results

        ii. Do not label the columns (your first row = first sample)
        iii. **Make sure to save as a .csv**
        iv. The majority of the columns will be empty
    b. Go to the **batch** GeoLocate client and upload the formatted csv file

c. "Page Georeference" will georeference all eight localities available at once. "Georeference" will do one at a time.
d. Select a locality and go through **Steps 1c** to **1g**. Once you are pleased with the locality and uncertainty click "Correct" to note that you have gone through this georeference.
e. Work through the remaining localities.
f. If you **do not** finish a batch georeferencing, you can click on "File Management" at the bottom of the screen to receive a retrieval code. This will allow you to re-access this file whenever you wish without the need to download and upload.
g. If you **do** finish a batch georeferencing, you can click on "File Management" and then "Export" to download the finished georeferenced file.

3. Use alternative resources to identify the localities in the example file. These are much more difficult and could use some historical maps and/or corrected spelling.

CLIMATE LAYER PROCESSING

QGIS provides a much better understanding of the processes happening with this step, but the R script streamlines a largely repetitive process. We will demo both options.

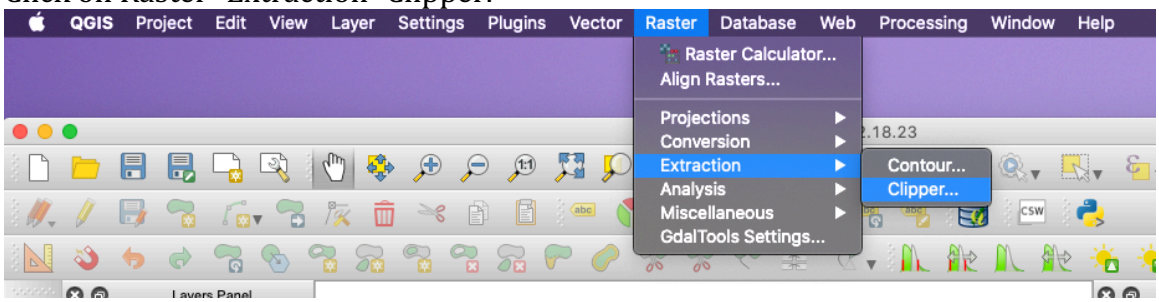
(A) Manual – QGIS - Optional

Files for this activity can be found in *"Demo/Manual/Climate_Layer_Processing/"* folder.

QGIS version has to be 2.18.23; if not, this will not work

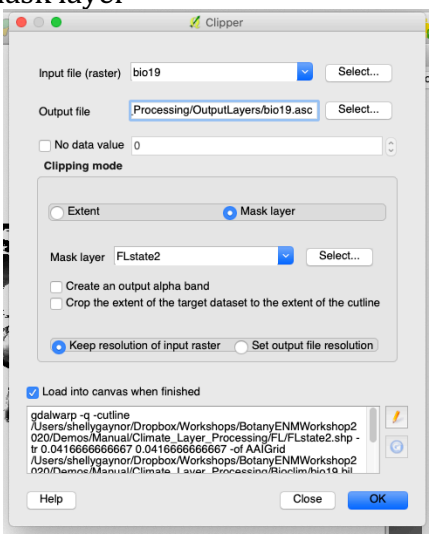
QGIS (version: 2.18.23)

1. Open QGIS.
2. Drag the layers (.bil files) found in the *"Demo/Manual/Climate_Layer_Processing/BioClim"* folder into QGIS. They should automatically appear. The box on the left lists the different layers not the layer is displayed. Drag *"Demo/Manual/Climate_Layer_Processing/FL/FLstate2.shp"* into QGIS and specify the CRS (Coordinate Reference System) as WGS 84.
3. Click on Raster>Extraction>Clipper.



4. Select the layer to clip in Input and designate an output folder and file name in Output. **Make sure the output file is in ASCII (.asc) format.**

5. Enter an Extent in the four latitude/longitude boxes or select an extent by moving the Clipping box and selecting directly on the background map. OR use Florida as a mask layer



6. Repeat **Step 4** for the remaining layers. Make sure not to change the extent.

(B) R based

- Open the R project by double clicking the .Rproj file. This can be found under *"Demo/Rbased/CrashCourse/CrashCourse.Rproj"*
    - Navigate to 03_*ClimateProcessing.R*
- Or follow along on the *CrashCourse_2020.html* file which can be found *"Demo/Rbased/CrashCourse_2020.html"*. This file can be opened in a web-browser.

CLIMATIC NICHE

(A) R based

- Open the R project by double clicking the .Rproj file. This can be found under *"Demo/Rbased/CrashCourse/CrashCourse.Rproj"*
    - Navigate to *04_PointBased.R*
- Or follow along on the *CrashCourse_2020.html* file which can be found *"Demo/Rbased/CrashCourse_2020.html"*. This file can be opened in a web-browser.

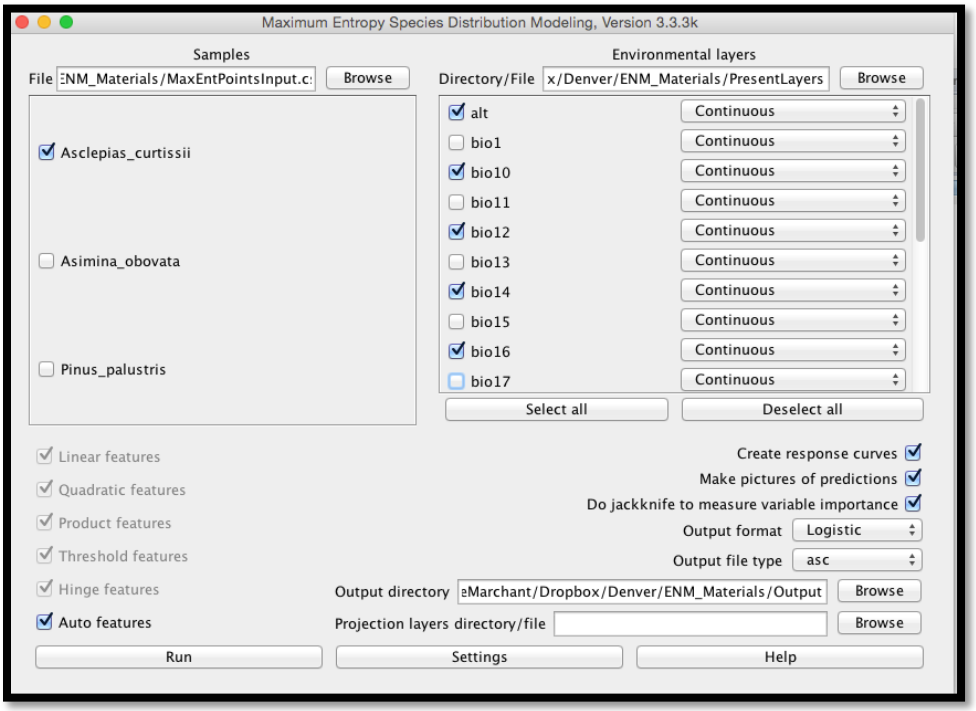
ECOLOGICAL NICHE MODELING

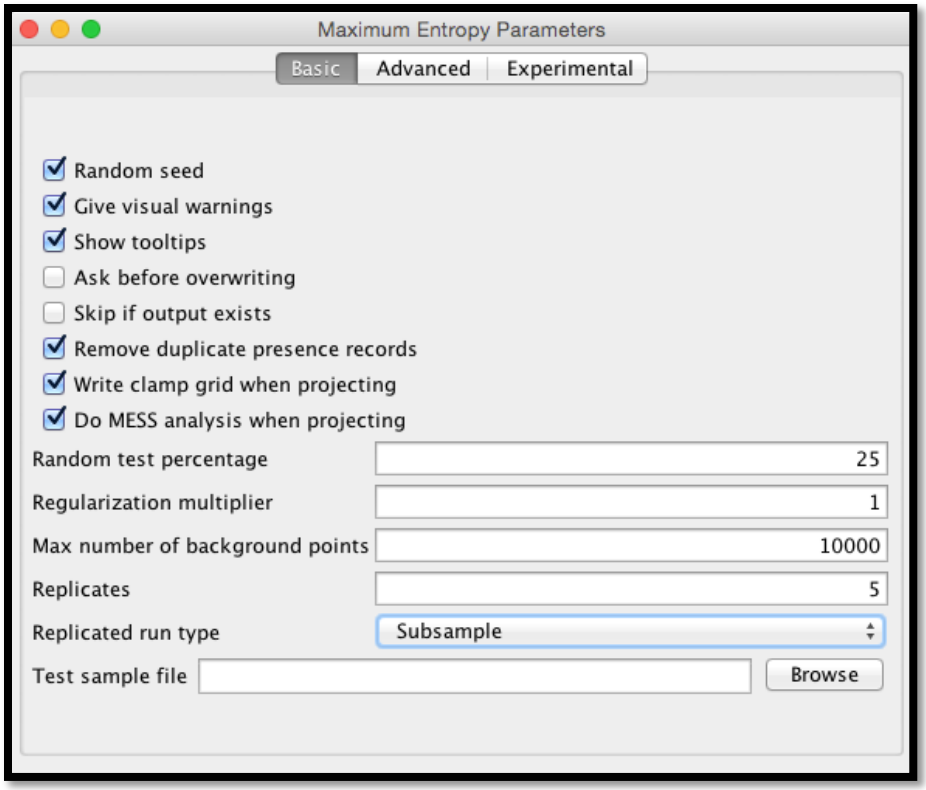
(A) Manual - MaxEnt

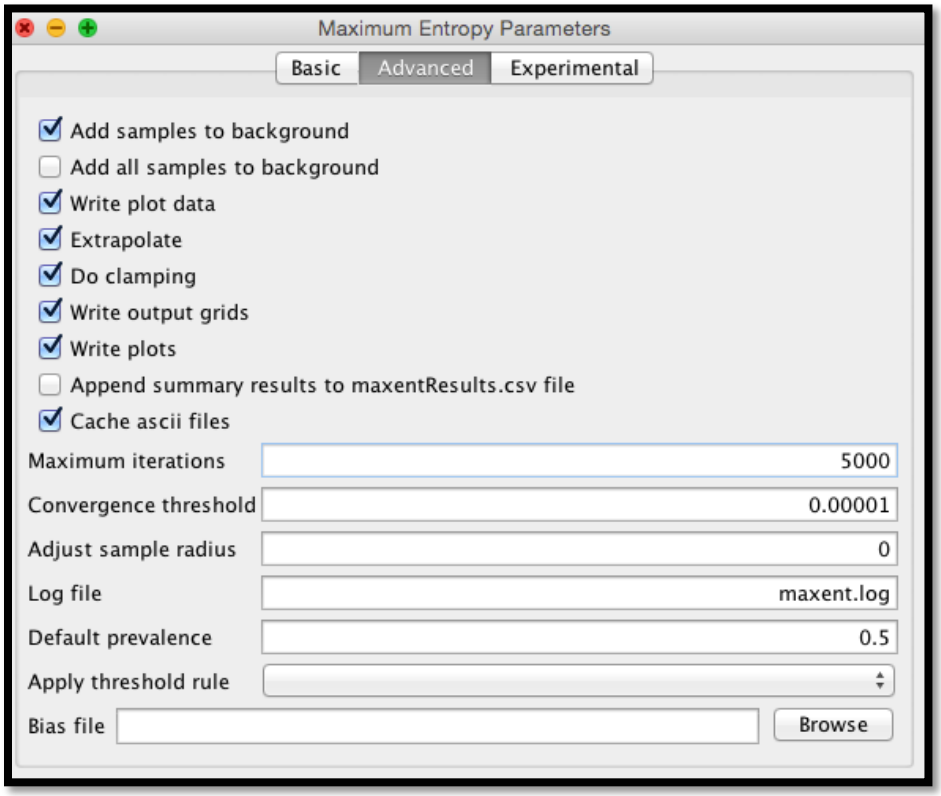
Files for this activity can be found in *"Demo/Manual/Ecological_Niche_Modeling/"* folder.

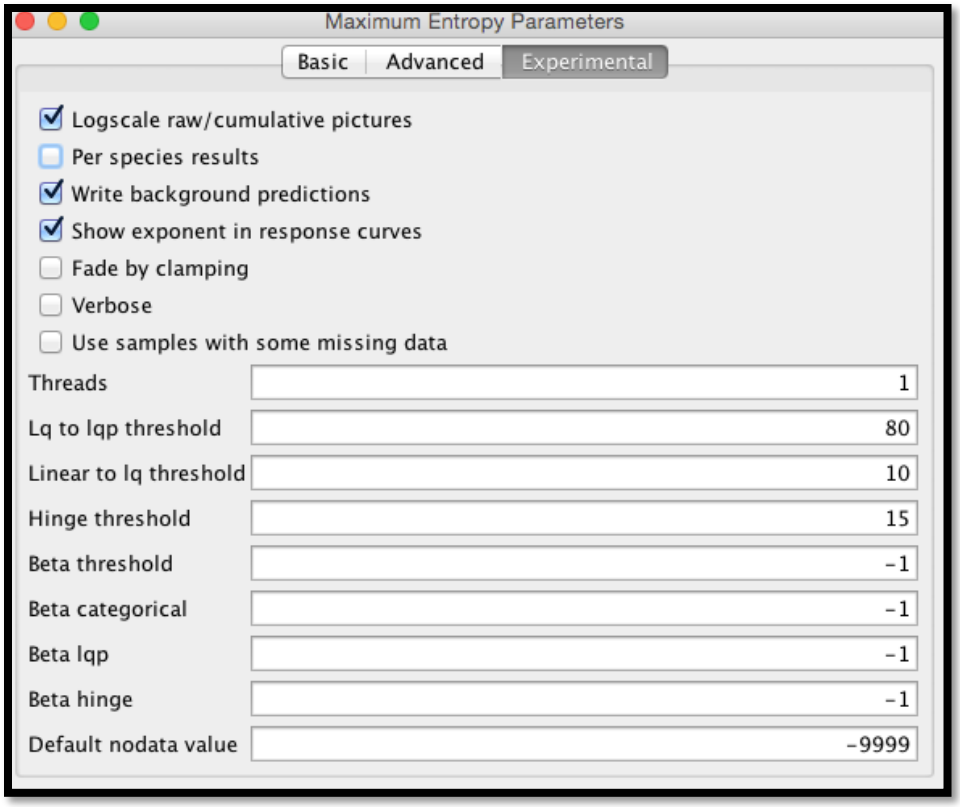
With our cleaned, georeferenced occurrence points and lowly-correlated, clipped layers, we are now ready to make some ecological niche models (ENMs).

1. Open Maxent (maxent.bar).
2. Select your cleaned occurrence csv file (MaxEntPointsInput.csv) in the Samples tab. Your species should become displayed in the box below. **Only run one species at a time.**
3. Select the folder with your Environmental layers (PresentLayers). All of the layers should become displayed, but only check the ones you selected from the correlation matrix.
4. Select an Output directory (Output).
5. We have attached screenshots of the parameters that should be selected and entered. Make sure to match yours with them.




Maximum Entropy Parameters
Basic
Advanced
Experimental
Random seed
Give visual warnings
Show tooltips
Ask before overwriting
Skip if output exists
Remove duplicate presence records
Write clamp grid when projecting
Do MESS analysis when projecting
Random test percentage
25
Regularization multiplier
1
Max number of background points
10000
Replicates
5
Replicated run type
Subsample
Test sample file
Browse


Maximum Entropy Parameters
Basic
Advanced
Experimental
Add samples to background
Add all samples to background
Write plot data
Extrapolate
Do clamping
Write output grids
Write plots
Append summary results to maxentResults.csv file
Cache ascii files
Maximum iterations
5000
Convergence threshold
0.00001
Adjust sample radius
0
Log file
maxent.log
Default prevalence
0.5
Apply threshold rule
Bias file
Browse



6. Click RUN!
7. If any errors pop up that says a point is missing environmental data, click "Ok"

(B) R based *Not included in today's workshop*

- Open the R project by double clicking the .Rproj file. This can be found under *"Demo/Rbased/CrashCourse/CrashCourse.Rproj"*
    - Navigate to *05_Ecological_Niche_Modeling.R*
- Or follow along on the *CrashCourse_2020.html* file which can be found *"Demo/Rbased/CrashCourse_2020.html"*. This file can be opened in a web-browser.

ECOLOGICAL NICHE MODEL PROCESSING

There are many additional analyses you may want to conduct after generating ENMs. This example is limited to only a few of those analysis.

(A) R based

- Open the R project by double clicking the .Rproj file. This can be found under "*Demo/Rbased/CrashCourse/CrashCourse.Rproj*"
    - Navigate to *06_ENM_Processing.R*

- Or follow along on the *CrashCourse_2020.html* file which can be found *"Demo/Rbased/CrashCourse_2020.html"*. This file can be opened in a web-browser.

PHYLOGENETIC DIVERSITY

(A) R based

- Open the R project by double clicking the .Rproj file. This can be found under "*Demo/Rbased/CrashCourse/CrashCourse.Rproj*"
    - Navigate to *07_Phylogenetic_Diversity.R*
- Or follow along on the *CrashCourse_2020.html* file which can be found at *"Demo/Rbased/CrashCourse_2020.html"*. This file can be opened in a web-browser.