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

9:00	Welcome and Overview of the Workshop – Pam
9:10	iDigBio Portal – Pam
9:20	Data Downloads – Lauren
9:35	Data Cleaning – Shelly
10:15	Outlier Detection – Tal
10:20	Georeferencing – Andre
10:50	Break
11:05	Climate Processing – Shelly
11:55	Climatic Niche – Shelly
12:15	Lunch
12:40	***Question Session***
1:00	Applications of ENMs – Doug
1:10	Ecological Niche Models – Shelly
1:40	Interpreting ENM Results – Shelly
2:00	Post-ENM analysis – Shelly
2:30	***Break/Question Session***
2:45	Intro to BiotaPhy – Doug
2:55	Overview of Phylogenetic Diversity (PD) – Hannah
3:10	Hands-on PD Demo – Maria
3:30	Overview of Alpine Biodiversity Project – Hector
3:40	Overview of OCBIL (Old, Climatically Buffered, Infertile Landscapes) Project – Maria
3:50	Correlating Chromosome and Climate Evolution – Jon
4:00	Computing with Heterogeneous Species Occurrence Data for Global Analyses – CJ
4:40	***Question Session***
5:00	End

Workshop Leaders:

Pam Soltis: psoltis@flmnh.ufl.edu
Doug Soltis: dsoltis@ufl.edu
Shelly Gaynor: michellegaynor@ufl.edu
Maria Cortez: mariacortez@ufl.edu
Andre Naranjo: aanaranjo@ufl.edu
Lauren Whitehurst: laurenwhitehurst@ufl.edu
Makenzie Mabry: mmabry44@gmail.com
Tal Kinser: tkinser@ufl.edu
Hector Figueroa: hecfox@umich.edu
Hannah Marx: marxh@umich.edu
Jon Spoelhof: spoelhof.jon@ufl.edu
Ryan Folk: rfolk@biology.msstate.edu
CJ Grady: cjgrady@ku.edu

SET-UP

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

(2) R and R Studio (demo built using R Versions 3.6.2)

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_Setup.R. Click on 00_Setup.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: 3.16)
    - MacOS: https://qgis.org/downloads/macos/qgis-macos-ltr.dmg
    - Windows: https://qgis.org/downloads/QGIS-OSGeo4W-3.16.8-4.msi
    - Other: https://qgis.org/en/site/forusers/download.html

iDigBio

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_2021.html* file which can be found *"Demo/Rbased/CrashCourse_2021.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. Decrease number of columns
3. Clean localities
    a. Rounds up the latitude/longitude to our desired coarseness and removes points that are not precise enough
    b. Removes coordinates at 0.00
    c. Removes coordinates in cultivated zones, botanical gardens, etc.
    d. Removes coordinates outside of our desired range
4. Removes duplicates
5. Spatial correction
6. Visualize
7. Produces a csv file with just the latitude and longitude for each record

(A)Manual – Optional

- Instructions are available in "Demos/Manual/Data_Cleaning/"
- Activity from 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.

(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_2021.html* file which can be found *"Demo/Rbased/CrashCourse_2021.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

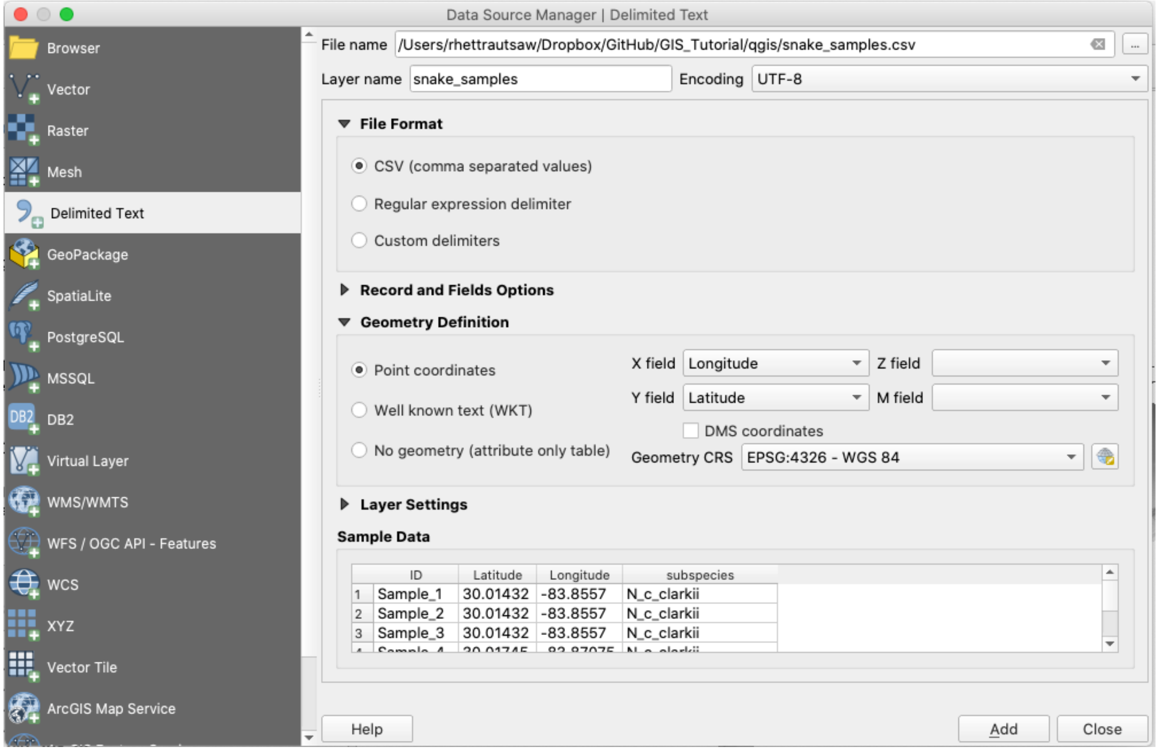
This activity was made by Rhett Rautsaw. Files for this activity can be found in *"Demo/Manual/Climate_Layer_Processing/"* folder.

QGIS version has to be 3.16; if not, this will not work

QGIS (version: 3.16)

1. Open QGIS.
2. Drag the layers (.tif files) found in the *"data/climate_processing/bioclim/"* folder into QGIS. They should automatically appear. The box on the left lists the different layers not the layer is displayed.
3. Add occurrence records from text-delimited file (Layer Menu > Add Layer > Add Delimited Text Layer…). Nagivate to "*data/cleaning_demo/maxent_ready/diapensiaceae_maxentready_20210625.csv*". X field is "longitude" and Y field is "latitude". Make sure the CRS is

EPSG:4326 – WGS 84.



File name /Users/rhettrautsaw/Dropbox/GitHub/GIS_Tutorial/qgis/snake_samples.csv

Layer name snake_samples Encoding UTF-8

▼ File Format

● CSV (comma separated values)

○ Regular expression delimiter

○ Custom delimiters

▶ Record and Fields Options

▼ Geometry Definition

● Point coordinates X field Longitude Z field

○ Well known text (WKT) Y field Latitude M field

○ No geometry (attribute only table) □ DMS coordinates

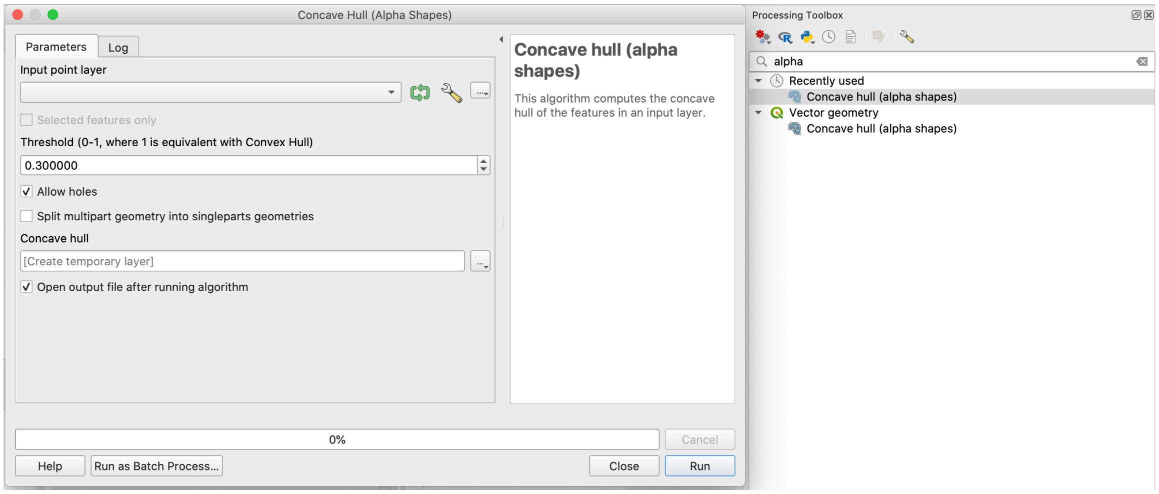
Geometry CRS EPSG:4326 - WGS 84

▶ Layer Settings

Sample Data

	ID	Latitude	Longitude	subspecies
1	Sample_1	30.01432	-83.8557	N_c_clarkii
2	Sample_2	30.01432	-83.8557	N_c_clarkii
3	Sample_3	30.01432	-83.8557	N_c_clarkii

4. Create an alpha hull/shape, using the Processing Toolbox Concave Hull Tool. Set the threshold to 1.



Input point layer

□ Selected features only

Threshold (0-1, where 1 is equivalent with Convex Hull)

0.300000

☑ Allow holes

□ Split multipart geometry into singleparts geometries

Concave hull

[Create temporary layer]

☑ Open output file after running algorithm

0%

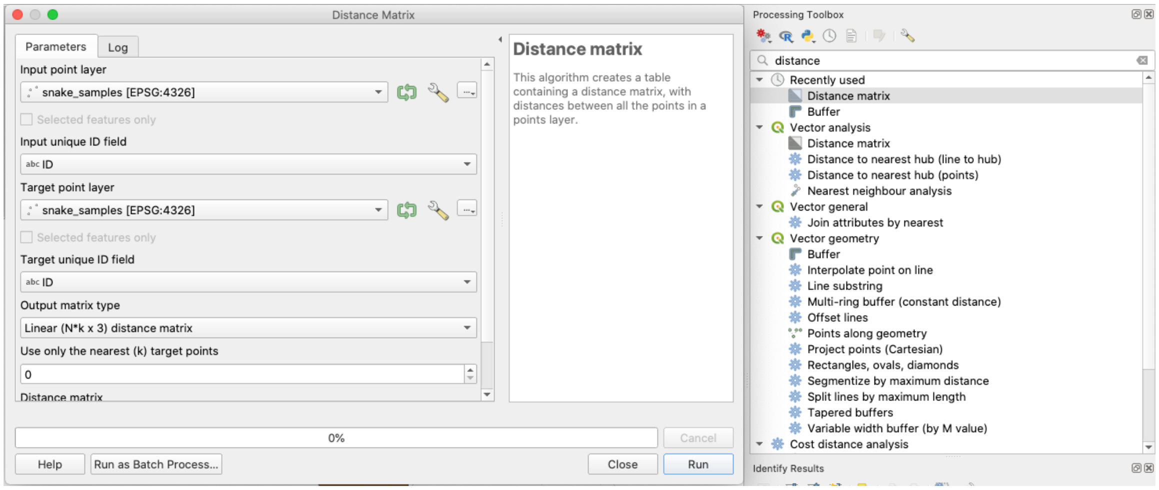
Concave hull (alpha shapes)

This algorithm computes the concave hull of the features in an input layer.

alpha

5. Calculate the greatest distance using the Processing Toolbox Distance Matrix Tool. Then open the Attributes Table for that matrix and use the last column to calculate the 80th quantile to find the suggested buffer

distance.



Distance Matrix

Parameters | Log

Input point layer: snake_samples [EPSG:4326]

□ Selected features only

Input unique ID field: abc ID

Target point layer: snake_samples [EPSG:4326]

□ Selected features only

Target unique ID field: abc ID

Output matrix type: Linear (N*k x 3) distance matrix

Use only the nearest (k) target points: 0

Distance matrix

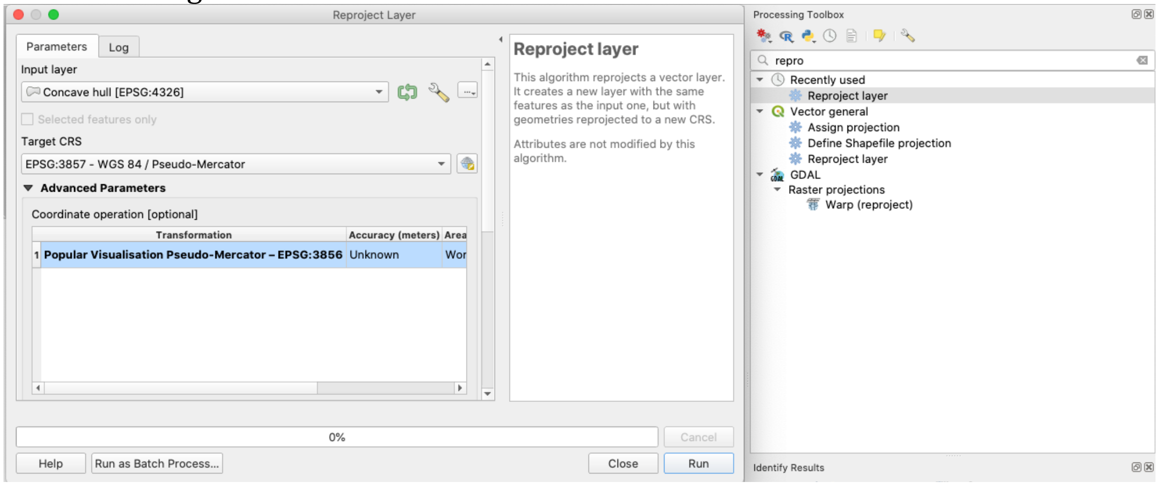
Distance matrix

This algorithm creates a table containing a distance matrix, with distances between all the points in a points layer.

0%

Cancel
Help
Run as Batch Process...
Close
Run
Processing Toolbox
distance
Recently used
Distance matrix
Buffer
Vector analysis
Distance matrix
Distance to nearest hub (line to hub)
Distance to nearest hub (points)
Nearest neighbour analysis
Vector general
Join attributes by nearest
Vector geometry
Buffer
Interpolate point on line
Line substring
Multi-ring buffer (constant distance)
Offset lines
Points along geometry
Project points (Cartesian)
Rectangles, ovals, diamonds
Segmentize by maximum distance
Split lines by maximum length
Tapered buffers
Variable width buffer (by M value)
Cost distance analysis
Identify Results

6. Reproject your alpha hull to a CRS in meters using the Reproject tool in the Processing Toolbox. Convert from EPSG:4326 to EPSG:3857.



Reproject Layer

Parameters | Log

Input layer: Concave hull [EPSG:4326]

□ Selected features only

Target CRS: EPSG:3857 - WGS 84 / Pseudo-Mercator

▼ Advanced Parameters

Coordinate operation [optional]

	Transformation	Accuracy (meters)	Area
1	Popular Visualisation Pseudo-Mercator – EPSG:3856	Unknown	Wor

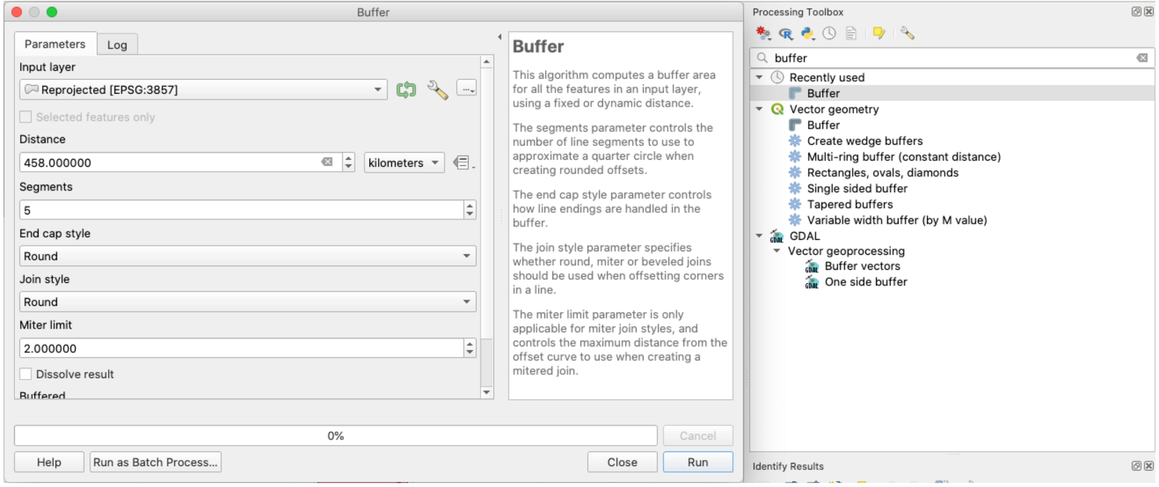
Reproject layer

This algorithm reprojects a vector layer. It creates a new layer with the same features as the input one, but with geometries reprojected to a new CRS.

Attributes are not modified by this algorithm.

0%

7. Buffer your reprojected layer by the suggested buffer distance using the Buffer tool in the Processing Toolbox.



Buffer

Input layer: Reprojected [EPSG:3857]

□ Selected features only

Distance: 458.000000 kilometers

Segments: 5

End cap style: Round

Join style: Round

Miter limit: 2.000000

□ Dissolve result

Buffered

Buffer

This algorithm computes a buffer area for all the features in an input layer, using a fixed or dynamic distance.

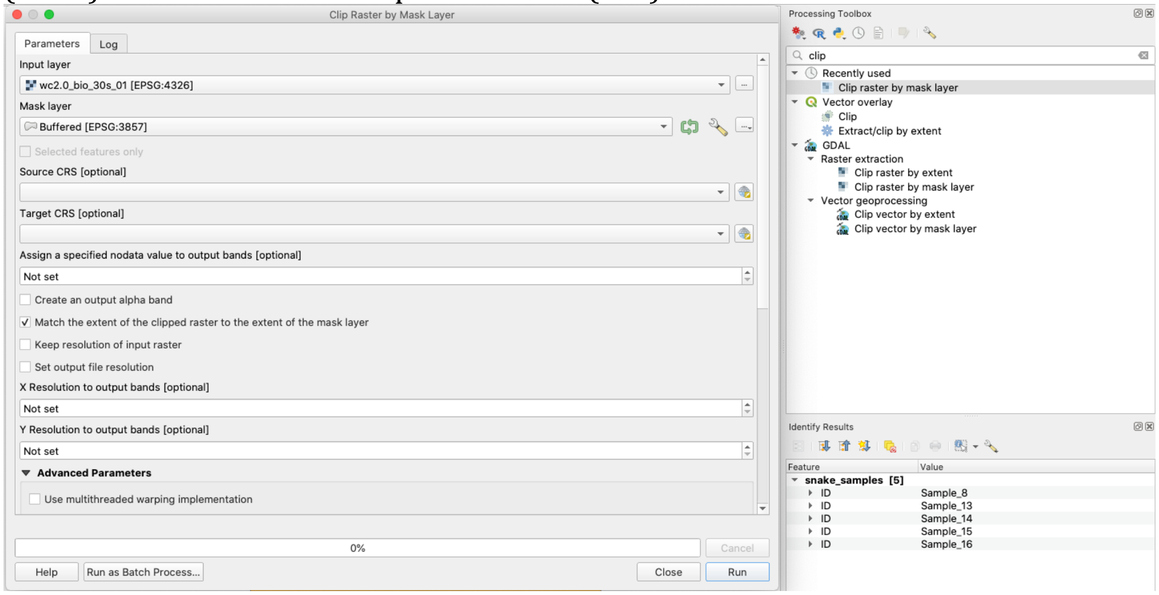
The segments parameter controls the number of line segments to use to approximate a quarter circle when creating rounded offsets.

The end cap style parameter controls how line endings are handled in the buffer.

The join style parameter specifies whether round, miter or beveled joins should be used when offsetting corners in a line.

The miter limit parameter is only applicable for miter join styles, and controls the maximum distance from the offset curve to use when creating a mitered join.

8. Next you can clip your rasters by your buffered layer using the "Clip raster by mask layer" in the Processing Toolbox. Scroll to the Clipped (mask) box and save this output to a ASCII (.asc) formatted raster.



Clip Raster by Mask Layer

Input layer: wc2.0_bio_30s_01 [EPSG:4326]

Mask layer: Buffered [EPSG:3857]

□ Selected features only

Source CRS [optional]

Target CRS [optional]

Assign a specified nodata value to output bands [optional]: Not set

□ Create an output alpha band

☑ Match the extent of the clipped raster to the extent of the mask layer

□ Keep resolution of input raster

□ Set output file resolution

X Resolution to output bands [optional]: Not set

Y Resolution to output bands [optional]: Not set

▼ Advanced Parameters

□ Use multithreaded warping implementation

Feature	Value
snake_samples [5]	
ID	Sample_8
ID	Sample_13
ID	Sample_14
ID	Sample_15
ID	Sample_16

Repeat for the remaining raster layers.

(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_2021.html* file which can be found *"Demo/Rbased/CrashCourse_2021.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_2021.html* file which can be found *"Demo/Rbased/CrashCourse_2021.html"*. This file can be opened in a web-browser.

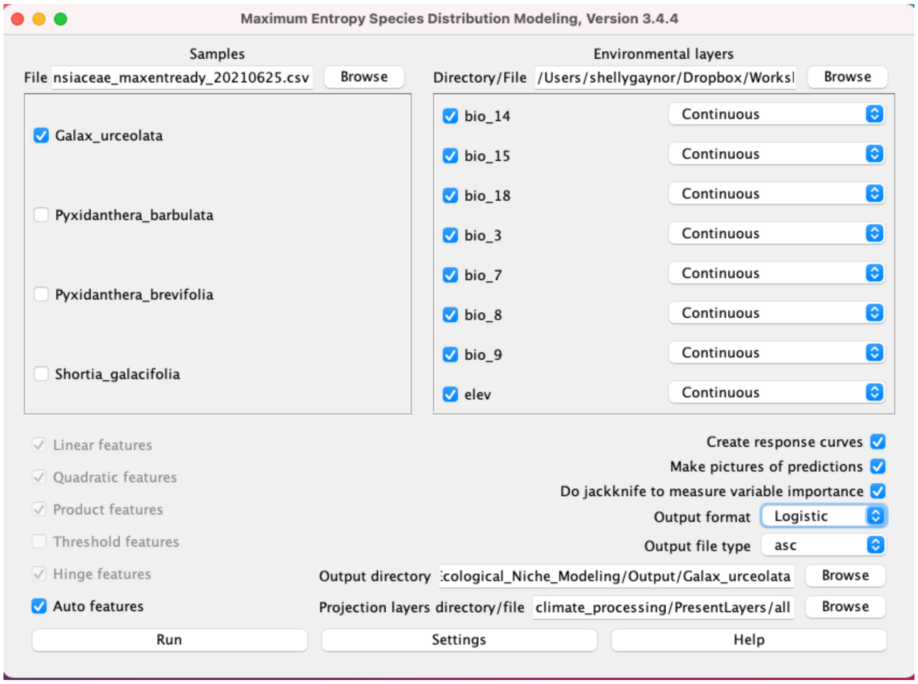
ECOLOGICAL NICHE MODELING

(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.jar).
    a. You may need to go the "System Preferences" -> "Security and Privacy" -> "General" and "Open anyways"
2. Select your cleaned occurrence csv file (diapensiaceae_maxentready_20210625.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 for one of your species (ex. "/Demos/Rbased/CrashCourse/data/climate_processing/PresentLayers/Galax_urceolata"). All of the layers should become displayed.
4. Add your Projected layers under "Projection layers directory/files" ("Demos/Rbased/CrashCourse/data/climate_processing/PresentLayers/all")
5. Select an Output directory (Output).
6. We have attached screenshots of the parameters that should be selected and entered. Make sure to match yours with them.



Maximum Entropy Species Distribution Modeling, Version 3.4.4

Samples

File nsiaceae_maxentready_20210625.csv

☑ Galax_urceolata

□ Pyxidanthera_barbulata

□ Pyxidanthera_brevifolia

□ Shortia_galacifolia

Environmental layers

Directory/File /Users/shellygaynor/Dropbox/Worksl

☑ bio_14	Continuous
☑ bio_15	Continuous
☑ bio_18	Continuous
☑ bio_3	Continuous
☑ bio_7	Continuous
☑ bio_8	Continuous
☑ bio_9	Continuous
☑ elev	Continuous

☑ Linear features

☑ Quadratic features

☑ Product features

□ Threshold features

☑ Hinge features

☑ Auto features

Create response curves ☑

Make pictures of predictions ☑

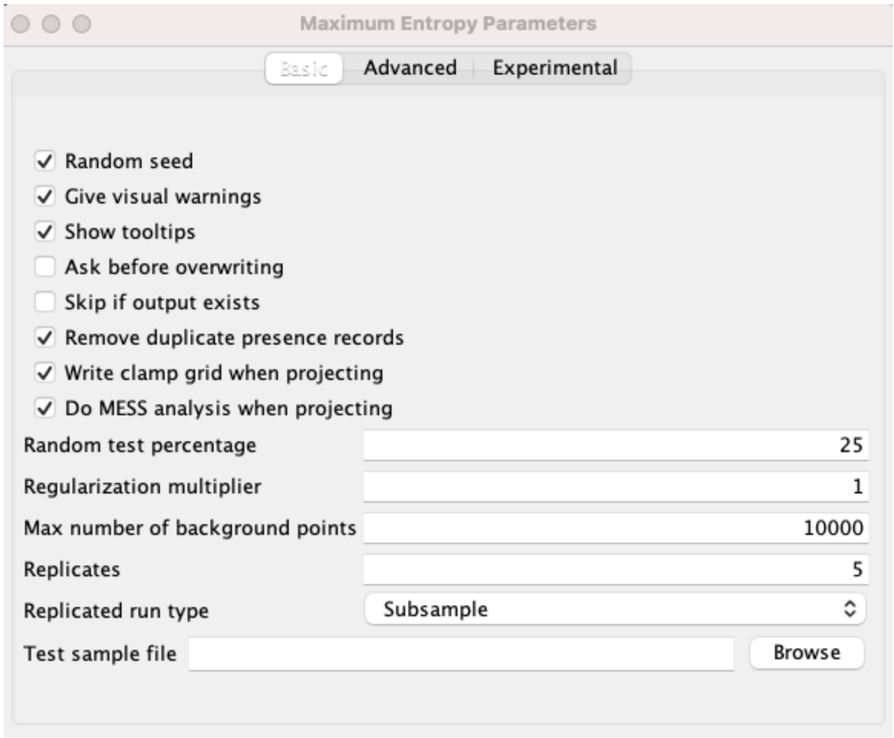
Do jackknife to measure variable importance ☑

Output format Logistic

Output file type asc

Output directory :cological_Niche_Modeling/Output/Galax_urceolata

Projection layers directory/file climate_processing/PresentLayers/all



Maximum Entropy Parameters

☑ Random seed

☑ Give visual warnings

☑ Show tooltips

□ Ask before overwriting

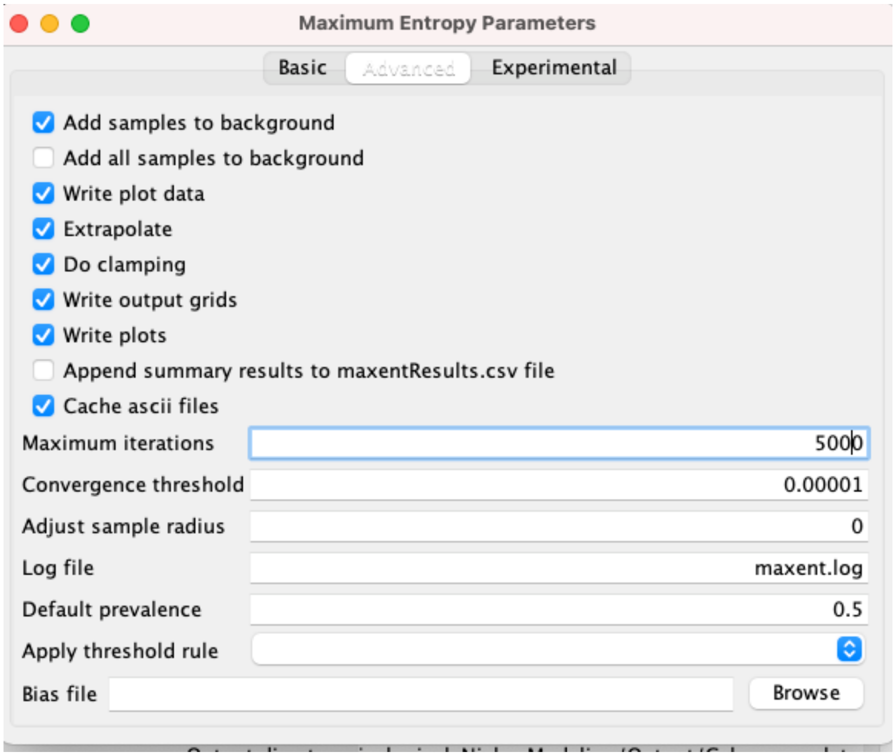
□ Skip if output exists

☑ Remove duplicate presence records

☑ Write clamp grid when projecting

☑ Do MESS analysis when projecting

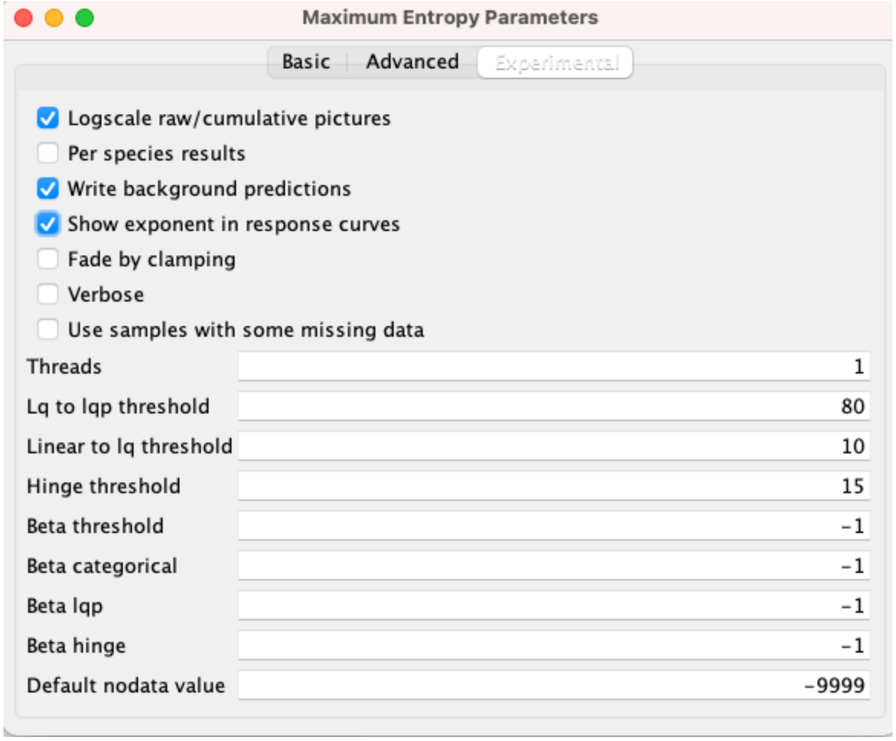
Random test percentage	25
Regularization multiplier	1
Max number of background points	10000
Replicates	5
Replicated run type	Subsample
Test sample file	



Maximum Entropy Parameters

☑ 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	



Maximum Entropy Parameters

☑ Logscale raw/cumulative pictures
☐ Per species results
☑ Write background predictions
☑ Show exponent in response curves
☐ Fade by clamping
☐ Verbose
☐ Use samples with some missing data

Threads	1
Lq to lqp threshold	80
Linear to lq threshold	10
Hinge threshold	15
Beta threshold	-1
Beta categorical	-1
Beta lqp	-1
Beta hinge	-1
Default nodata value	-9999

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

(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 *05_Ecological_Niche_Modeling.R*
- Or follow along on the *CrashCourse_2021.html* file which can be found *"Demo/Rbased/CrashCourse_2021.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_2021.html* file which can be found *"Demo/Rbased/CrashCourse_2021.html"*. This file can be opened in a web-browser.

BIOTAPHY

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_2021.html* file which can be found at *“Demo/Rbased/CrashCourse_2021.html”*. This file can be opened in a web-browser.