RANGE AREA MODELING DECISION TREE

The following decision tree will be used to determine how each species is modeled so that we are able to calculate a range area for each species.

A. >=5 unique [1] points………………………………………………………… Model using Maxent
A. <5 unique points, or failed Maxent…………………………………… B

B. 3-4 unique points or failed Maxent ………………………………… Three-Four point algorithm
B. <3 unique points………………………………………………………………. C

C. 2 unique points……………………………………………………………….. Two point algorithm
C. 1 unique point………………………………………………………………… Single point algorithm

Three-Four point algorithm:

X or Y values equal……………………………………………………………… Fit all points as 1-point boxes [3]
X or Y values distinct…………………………………………………………… Fit convex hull

Two point algorithm:

Create one box around both points or a separate box around each point?
Calculate a bounding box around both points and compare area of box with Gentry area.

X or Y values equal……………………………………………………………… Fit as two 1-point boxes [3]
Bounding Box area > Gentry area (75,000 km^2)…………………… Fit as two 1-point boxes [2]
Bounding Box area <= Gentry area (75,000 km^2)…………………. Fit single box [3]

One point algorithm:

Calculate one Gentry box around each point [3]

Notes:

[1] Unique defined as having distinctly separate geographic locations (lat/long)
[2] Two point boxes will be constructed using the bounding latitudes and longitudes of both points.
[3] Single point boxes will be constructed by setting the point as the center of the box and extending 136.9 km in each direction to create a box of area approx. 75,000 km2

Gentry cutoff for small range species:

"We defined small-ranged species as those that have never been collected outside the Peruvian department of Madre de Dios, and large-ranged species as those that have. The rationale is that the size of the department of Madre de Dios, 78,415 km2, fits the most widely used criterion of tropical plant species endemism, 50–75 000 km2 (Gentry 1986, 1992)." - From Pitman 1999

Pitman, N. C. a., Terborgh, J., Silman, M. R., & Nuñez V., P. (1999). Tree Species Distributions in an Upper Amazonian Forest. Ecology, 80(8), 2651-2661. doi:10.1890/0012-9658(1999)080[2651:TSDIAU]2.0.CO;2Gentry. 1986. Endemism in tropical vs. temperate plant communities. Pages 153–181 in M. E. Soule´, editor. Conservation biology: the science of scarcity and diversity. Sinauer Associates, Sunderland, Massachusetts, USA.

Gentry. 1988. Changes in plant community diversity and floristic composition on environmental and geographical gradients. Annals of the Missouri Botanical Garden 75(1): 1–34.

DATA PRODUCTS WRITTEN PER SPECIES

The following describes the data products, formats and projections that are written for each range model based on the number of geographically distinct occurrences for each species.

Data Product Files Output	Format	1 Pt	2 Pts	3 Pts	4 Pts	5+ Pts
Occurrence Points	CSV	Y	Y	Y	Y	Y
Gentry Box	Shapefile	UTM WGS	UTM WGS			
Convex Hull (Clipped)	Shapefile			UTM WGS	UTM WGS	UTM WGS
Bounding Box (Clipped)	Shapefile			UTM WGS	UTM WGS	UTM WGS
Latitude Extent (Clipped Band)	Shapefile			UTM WGS	UTM WGS	UTM WGS
Occupied Cells (raster points)	ERDAS GeoTiff	UTM	UTM	UTM	UTM	UTM
Maxent Logistic Surface	ERDAS GeoTiff					UTM WGS
Maxent Binary Surface (Raster)	ERDAS GeoTiff					UTM
Maxent Binary Map (Polygon)	Shapefile					UTM WGS
Recorded Statistics	CSV					
Species Name	.	Y	Y	Y	Y	Y
Number of Samples	.	Y	Y	Y	Y	Y
Number of Unique Points	.	Y	Y	Y	Y	Y
Gentry Box Type [*]	.	Y	Y			
Gentry Box Area [1]	.	Y	Y			
Convex Hull Area Unclipped [1]	.			Y	Y	Y
Bounding Box Area Unclipped [1]	.			Y	Y	Y
Latitudinal Range (Linear) [2]	.		Y	Y	Y	Y
Longitude Range (Linear) [2]	.		Y	Y	Y	Y
Latitudinal Extent (Unclipped Band) [1]	.			Y	Y	Y
Convex Hull Area Clipped [1]	.			Y	Y	Y
Bounding Box Area Clipped [1]	.			Y	Y	Y
Latitudinal Extent (Clipped Band) [1]	.			Y	Y	Y
Cell Size [3]	.	Y	Y	Y	Y	Y
Area of Occupancy (Cell area) [3]	.	Y	Y	Y	Y	Y
Maxent Areas (3 models) [3,4]	.					Y
Maxent All Thresholds	CSV					Y
MaxentResults.csv	CSV					Y

[*] Describes how the Gentry Area was defined (by a single Gentry box, two boxes, etc.)

[1] Area calculated from the geometric shape – not a raster shape

[2] Area calculated by subtracting max from min lat/long coordinates

[3] Area calculated from a raster image and based on number of cells * cell size

[4] Areas from the three Maxent model types (bioclim, spatial filters, bioclim + spatial filters) and various thresholds defined below:

- Maxent minimum training presence threshold area
- Maxent 1 percent training presence threshold area

- Maxent 5 percent training presence threshold area
- Maxent max Kappa threshold area
- Maxent balanced sensitivity and specificity area
- Maxent maximum sensitivity and specificity area

DATA PRODUCT SIZE ESTIMATES

The following are some rough size estimates for the files written by each model type. All sizes are expressed in kilobytes.

Model Products	**Format**	**1 Pt**	**2 Pts**	**3-4 Pts**	**5+ Pts**
Gentry Box	Shapefiles	3	3	-	-
Convex Hull (Clipped)	Shapefiles	-	-	2	2
Bounding Box (Clipped)	Shapefiles	-	6	6	6
Latitude Extent (Clipped Band)	Shapefiles	-	8	8	8
Occupied Cells (raster points)	Rasters	6,000	6,000	6,000	6,000
Maxent Logistic Surface	Rasters	-	-	-	27,000
Maxent Binary Surface (Raster)	Rasters	-	-	-	6,000
Maxent Binary Map (Polygon)	Shapefiles	-	-	-	128
Recorded Statistics					
Base Statistics	CSV	325	365	420	420
Maxent Area Statistics	CSV	-	-	-	985
Maxent All Thresholds	CSV	-	-	-	34
MaxentResults	Folder	-	-	-	11,000
Size per Model		6,328	6,382	6,436	51,583
Total Number of Species		20,680	10,766	46,135	46,135
Total Size Needs Per Model		130,863,040	68,708,612	296,924,860	2,379,781,705

Grand Total Size Needs: 2,876,278,217 kb (or approximately 2.68 TB)

BIEN MODELING SCRIPTS

The following scripts are used in modeling range areas for the BIEN database. Scripts which have a Cmd value of "Yes" indicates scripts that are written to be run as commands from the Linux command line.

Script File Name	Description	Cmd
AnyPointAlgorithm.r	Reads a CSV file containing species occurrence records and determines which specific modeling algorithm to run based on the number of geographically distinct points. Calls SinglePointAlgorithm, TwoPointAlgorithm	Yes
SinglePointAlgorithm.r	Called by AnyPointAlgorithm. Reads a CSV file containing 1 species occurrence record and generates a Gentry box and records area statistics.	
TwoPointAlgorithm.r	Called by AnyPointAlgorithm. Reads a CSV file containing 2 species occurrence records and determines which specific 2-point modeling algorithm to run. Calls SinglePointAlgorithm, TwoPointEvlauationBoundingBox, TwoPointSingleBoundingBox, MakeSeparateGentryBoxes	
TwoPointEvaluationBoundingBox.r	Called by TwoPointAlgorithm. Creates a bounding box from two species occurrence points and uses the resulting box area to determine whether range area should be calculated as a single bounding box or the sum of separate Gentry boxes. Returns box type to calculate.	
TwoPointSingleBoundingBox.r	Called by TwoPointAlgorithm. Creates a bounding box from two species occurrence points and records area statistics.	
MakeSeparateGentryBoxes.r	Called by TwoPointAlgorithm. Creates separate Gentry boxes for each supplied point, sums the area of each box and records area statistics.	
MakeMultipleRangeAreas.r	Called by AnyPointAlgorithm (if number of occurrence points > 2). Generates bounding box, convex hull, latitudinal band, latitudinal extent, longitudinal extent, and cell occupancy area (and linear distance) measurements and records area statistics.	
CreateSDM_Bioclim.r	Called by NONE. Reads a CSV file containing at least 5 geographically distinct species occurrence points and generates a Maxent species distribution model using 19 bioclim environmental layers. Records raw Maxent model, area statistics resulting from several thresholds, and model accuracy values. Saves binary surface and polygon of model resulting from 1% training presence threshold that has been specially clipped. Also records all thresholds in a separate file.	Yes
CreateSDM_BioclimSpatial.r	Called by NONE. Reads a CSV file containing at least 5 geographically distinct species occurrence points and generates a Maxent species distribution model using 19 bioclim and 19 Spatial Filter environmental layers. Records raw Maxent model, area statistics resulting from several thresholds, and model accuracy values. Saves binary surface and polygon of model resulting from 1% training presence threshold that has been specially clipped. Also records all thresholds in a separate file.	Yes
CreateSDM_Spatial.r	Called by NONE. Reads a CSV file containing at least 5 geographically distinct species occurrence points and generates a Maxent species distribution model using 19 Spatial Filter environmental layers. Records raw Maxent model, area statistics	Yes

	resulting from several thresholds, and model accuracy values. Saves binary surface and polygon of model resulting from 1% training presence threshold that has been specially clipped. Also records all thresholds in a separate file.	
gdal_polygonize.py	Called by CreateSDM_Bioclim/Spatial/BioclimSpatial models. Generates a polygon shapefile of from a binary Maxent raster model in raster format.	Yes

EXAMPLE COMMANDS

The following are examples of how to execute the AnyPointAlgorithm command and the CreateSDM_Bioclim, CreateSDM_BioclimSpatial and CreateSDM_Spatial commands in the Linux command line.

- **AnyPointAlgorithm**

```
R --slave --args
"/data/project/bien/John/BIENRangeModelingFinal/SpeciesCSVs/BIENAll/UTM/1Point/Aa_f
iebrigii.csv" "/data/project/bien/John/BIENRangeModelingFinal"
"/data/project/bien/John/BIENRangeModelingFinal/BackgroundRaster/woGreenland/backgr
ound.img"
"/data/project/bien/John/BIENRangeModelingFinal/BackgroundShapefiles/woGreenland/Ne
wWorld.shp" "/data/project/bien/John/BIENRangeModelingFinal/Output" <
AnyPointAlgorithm.r
```

Arguments:
<Path to species occurrence CSV file>
<Working Directory>
<Path to Background Raster>
<Path to background Shapefile>
<Output directory> -- *Directory must previously exist*

- **CreateSDM_Bioclim**

```
R --slave --args
"/data/project/bien/John/BIENRangeModelingFinal/SpeciesCSVs/BIENAll/UTM/5Points/Pin
us_edulis.csv" "/data/project/bien/John/BIENRangeModelingFinal"
"/data/project/bien/John/BIENRangeModelingFinal/Output" FALSE < CreateSDM_Bioclim.r
```

Arguments:
<Path to species occurrence CSV file>
<Working Directory>
<Output directory> -- *Directory must previously exist*

Assumes Background Raster and environmental layers are in the following subdirectories under the working directory:
/BackgroundRaster/woGreenland/background.img
/EnvironmentalRasters/NewWorld/woGreenland/BioClim1-19/.asc*
/EnvironmentalRasters/NewWorld/woGreenland/Spatial-19/.asc*

- **CreateSDM_Spatial**

```
R --slave --args
"/data/project/bien/John/BIENRangeModelingFinal/SpeciesCSVs/BIENAll/UTM/5Points/Pin
us_edulis.csv" "/data/project/bien/John/BIENRangeModelingFinal"
"/data/project/bien/John/BIENRangeModelingFinal/Output" FALSE < CreateSDM_Spatial.r
```

Arguments:
Same as above

- **CreateSDM_Bioclim_Spatial**

```
R --slave --args
"/data/project/bien/John/BIENRangeModelingFinal/SpeciesCSVs/BIENAll/UTM/5Points/Pin
us_edulis.csv" "/data/project/bien/John/BIENRangeModelingFinal"
"/data/project/bien/John/BIENRangeModelingFinal/Output" FALSE <
CreateSDM_BioclimSpatial.r
```

Arguments:

Same as above

MODEL RUN TIMES

The following are estimates for per-species model run times.

Model	Run Time (in sec)
SinglePointAlgorithm.r	12 *
TwoPointSingleBoundingBox.r	14 *
MakeSeparateGentryBoxes.r	16 *
MakeMultipleRangeAreas.r	25 *
CreateSDM_Bioclim.r	840
CreateSDM_BioclimSpatial.r	1,020
CreateSDM_Spatial.r	840
gdal_polygonize.py	7 *

* *The run-time was manually estimated; precise metrics were not recorded*

DIRECTORY STRUCTURE

The following is the directory structure necessary for a range model run. File names have been simplified for easier reading.

Base Directory Structure

```
/Working Directory
    /BackgroundRaster/woGreenland/
        background.img
    /BackgroundShapefiles/woGreenland/
        NewWorld.shp
    /EnvironmentalRasters/NewWorld/woGreenland/BioClim1-19/
        bio_*_p.asc (* = 1-19)
    /EnvironmentalRasters/NewWorld/woGreenland/Spatial1-19/
        filter*_10p.asc (* = 001 to 019)
    /Output/
        subdirectories below the output directory will be automatically created by the scripts
    /Scripts/
        scripts in this directory are documented above
    /SpeciesCSVs/BIENAll/UTM/
        *.csv – species occurrences named as genus_species.csv
```

Output Subdirectory Structure

The model scripts will automatically create subdirectories below the output directory to store model output. The following subdirectories are created.

/**Output**

- /**BoundingBox**
  - /**UTM** – *Stores shapefile versions of bounding box range area*
  - /**WGS** – *Stores shapefile versions of bounding box range area*
- /**ConvexHull**
  - /**UTM** – *Stores shapefile versions of convex hull range area*
  - /**WGS** – *Stores shapefile versions of convex hull range area*
- /**Gentry**
  - /**UTM** – *Stores shapefile versions of Gentry box range area(s)*
  - /**WGS** – *Stores shapefile versions of Gentry box range area(s)*
- /**LatExtent**
  - /**UTM** – *Stores shapefile versions of latitudinal band range area*
  - /**WGS** – *Stores shapefile versions of latitudinal band range area*
- /**Maxent**
  - /**Maxent_Bioclim**
    - /**ERDAS**
      - /**UTM** – *Stores raster versions of Maxent raw model & binary range area*
      - /**WGS** – *Stores raster versions of Maxent raw model & binary range area*
    - /**GeoTiff**
      - /**UTM** – *Stores raster versions of Maxent raw model & binary range area*
      - /**WGS** – *Stores raster versions of Maxent raw model & binary range area*
    - /**Shapefile**
      - /**UTM** – *Stores shapefile versions of Maxent binary range area*

- /**WGS** – *Stores shapefile versions of Maxent binary range area*
  - /**Maxent_BioclimSpatial**
    - *Same as above*
  - /**Maxent_Spatial**
    - *Same as above*
- /**Points**
  - /**UTM** – *Stores raster versions of area of occupancy range areas*
- /**Statistics**
  - /**Base** – *Stores basic statistics of calculated range areas*
  - /**Maxent_Bioclim** – *Stores statistics of Maxent model range areas*
  - /**Maxent_BioclimSpatial** – *Stores statistics of Maxent model range areas*
  - /**Maxent_Spatial** – *Stores statistics of Maxent model range areas*
  - /**Thresholds**
    - /**Maxent_Bioclim** – *Stores all thresholds saved from Maxent models*
    - /**Maxent_BioclimSpatial** – *Stores all thresholds saved from Maxent models*
    - /**Maxent_Spatial** – *Stores all thresholds saved from Maxent models*

MAXENT TEMP FILE STORAGE

Running the Maxent models generates temporary files for each model run. When run on the cluster, these temp files are written by species. We need to store them so they can be mined for additional information on variable importance and other subsequent analyses.

Each species will generate approximately 10.5 MB of temp files
Directory structure is as follows:

- /temp/genus_species/
    - genus_species.asc
    - genus_species.html
    - genus_species.lambdas
    - genus_species_omission.csv
    - genus_species_sampleAverages.csv
    - genus_species_samplePredictions.csv
    - maxent.log
    - maxentResults.csv
    - /plots/
        - *Between 80 and 156 PNG image files*

Example:

- /temp/Acacia_greggii/
    - Acacia_greggii.asc
    - Acacia_greggii.html
    - Acacia_greggii.lambdas
    - Acacia_greggii _omission.csv
    - Acacia_greggii _sampleAverages.csv
    - Acacia_greggii _samplePredictions.csv
    - maxent.log
    - maxentResults.csv
    - /plots/
        - *Between 80 and 156 PNG image files*

While it would be advantageous to save the entire temp folder for each species, at a minimum we need to preserve the file named maxentResults.csv

The file will need to be copied from the /temp/genus_species/ folder location and renamed from maxentResults.csv to genus_species_maxentResults.csv and stored with the rest of the model output such as within a folder inside of the /Statistics/<Maxent Model Type>/ folder. For example:

```
/Statistics/Maxent_Bioclim/MaxentResults/genus_species_maxentResults.csv
```

R-ENVIRONMENT AND PACKAGE VERSIONS

The following R configuration is known to work well for running the range modeling scripts in the R environment on Linux.

- **R version**
  - R version 2.15.0 (2012-03-30)
  - Platform: x86_64-pc-linux-gnu (64-bit)
- Package versions: items in **bold** are specifically loaded by range modeling scripts.
  - **rgeos_0.2-6**
  - plyr_1.7.1
  - stringr_0.6
  - **maptools_0.8-14**
  - lattice_0.20-6
  - foreign_0.8-50
  - **PBSmapping_2.62.34**
  - **igraph_0.5.5-4**
  - **dismo_0.7-17**
  - **raster_1.9-92**
  - **rJava_0.9-3**
  - **rgdal_0.7-8**
  - **sp_0.9-99**
- Packages loaded via a namespace (and not attached)
  - grid_2.15.0
  - lattice_0.20-6
  - tools_2.15.0

- End -