

Ecological niche modeling

Data access and use workshop

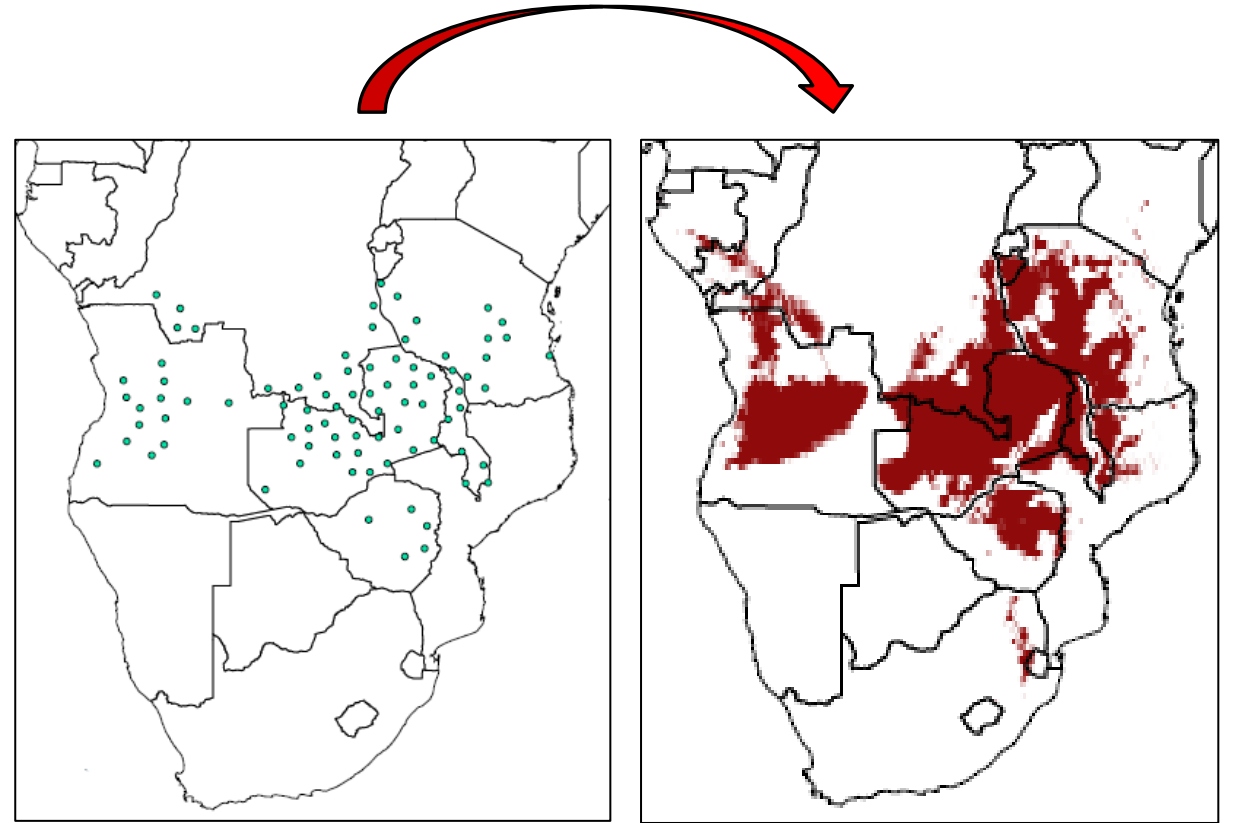
March 2020

Bindura University of Science Education

Percy Jinga and Admore Mureva

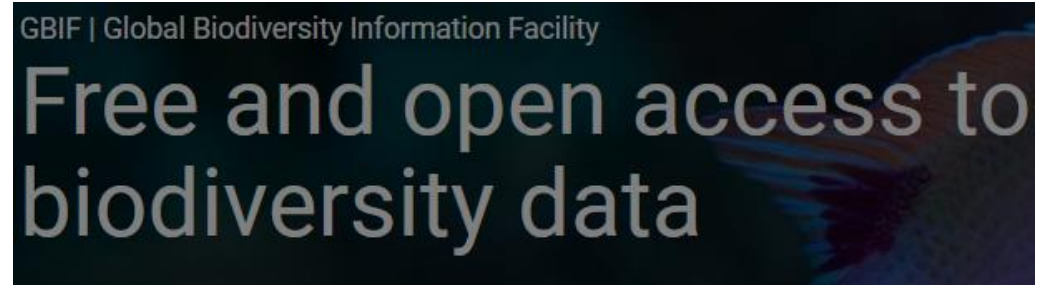
Ecological niche models (ENMs)

- Predictive modeling of species geographic distributions based on environmental conditions at sites of known occurrence
- Applications in conservation, reserve planning, ecology, epidemiology and invasive-species management



Occurrence records

- Each occurrence locality is a latitude-longitude pair denoting a site where the species has been observed
- Georeferenced occurrence records are derived from specimens in natural history museums and herbaria as well as online data portals, such as GBIF (www.gbif.org), Tropicos (www.tropicos.org) and Species Link (<http://splink.cria.org.br/>)



Environmental variables

- Environmental variables, in GIS format, are from the study area partitioned into a grid of pixels
- Examples include climate variables, altitude, vegetation cover and soil characteristics
- Commonly obtained from online data portals, such as WorldClim (www.worldclim.org/) and Harmonized World Soil Database (www.fao.org/)

WorldClim - Global Climate Data

Free climate data for ecological modeling and GIS



Food and Agriculture Organization
of the United Nations

FAO SOILS PORTAL



Survey

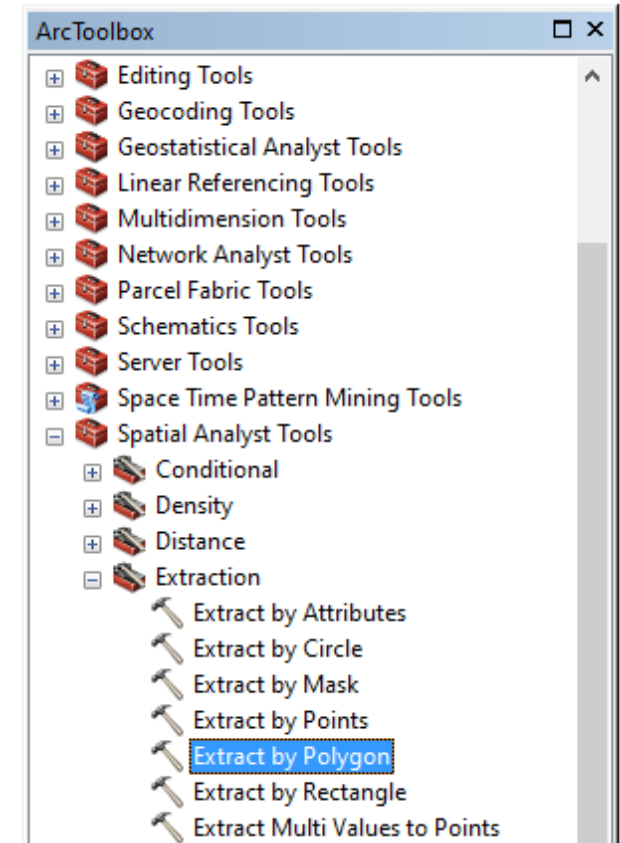
Assessment

Biodiversity

Management

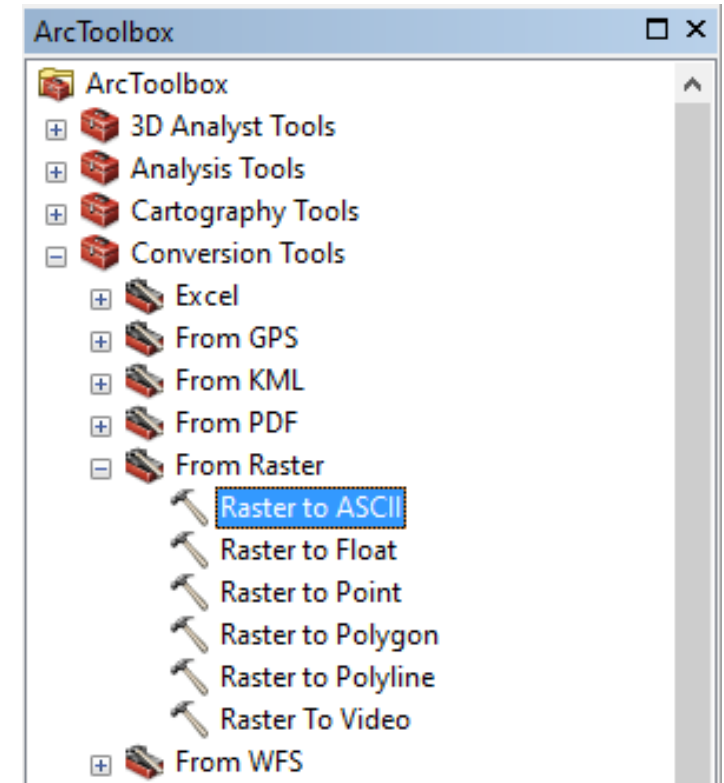
Preparation of environmental layers

- Extract files to restrict modeling to the area of interest
- Hover the mouse in AcrMap to get coordinates of interest
- Use the coordinates to “Extract by Polygon.”
- Make sure to close the polygon, the starting pair of coordinates should be the last



Preparation of environmental layers (cont'd)

- Convert all extracted files from raster to Ascii
- Make sure the “Extent” and “cell size” are identical for all files



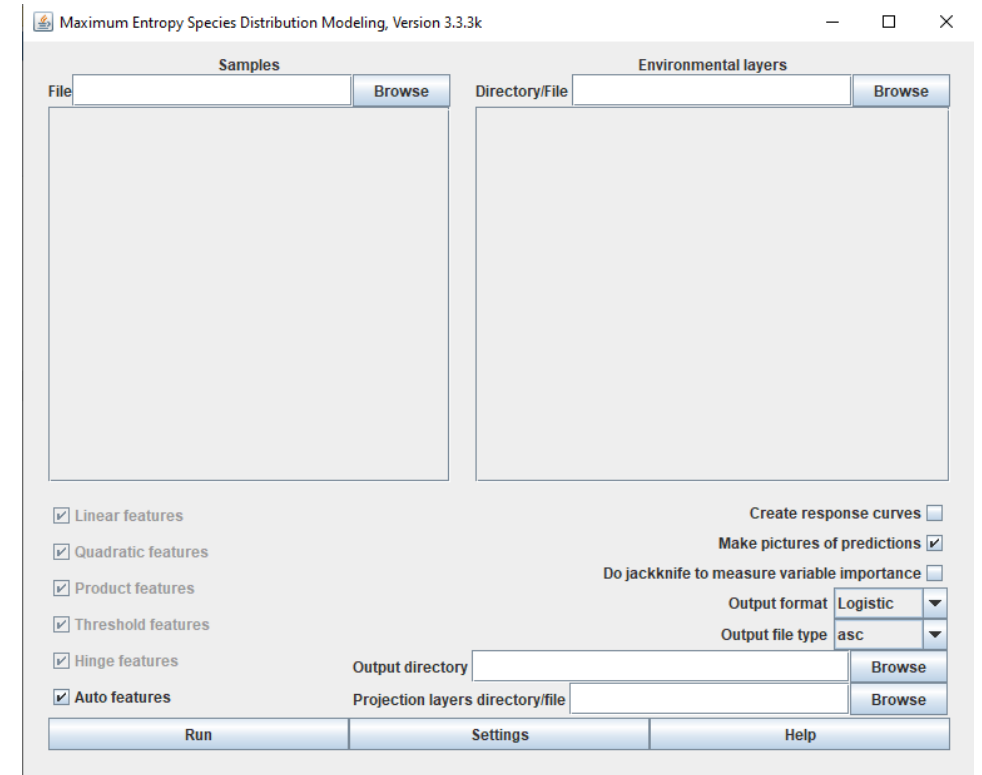
Modeling

- A niche-based model represents an estimated species' ecological niche in the examined environmental dimensions
- Although the model describes suitability in ecological space, it is typically projected into geographic space, yielding a geographic area of predicted presence
- Geographic barriers and biotic interactions entails few species occupy all areas with optimum niche requirements



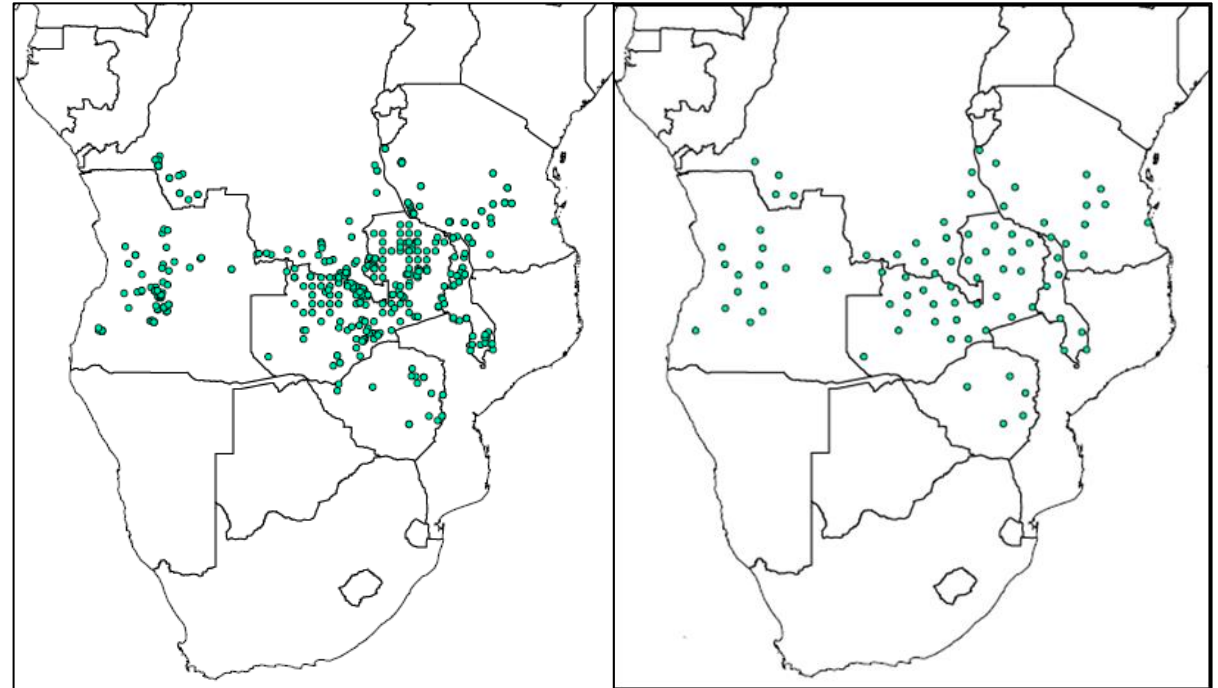
Maximum entropy method (MaxEnt)

- Use presence-only data, a significant advantage since absence data is not commonly available in poorly sampled areas
 - Better results with small number of occurrence records
 - High predictive accuracy compared to other methods
 - User-friendly graphical user interface
- Can be run in R



Resampling occurrence records

- Records may be biased; may be obtained from easily accessible areas, such as near roads and human settlements
- Different sampling intensity across the study landscape
- Predictive accuracy will be affected because of overrepresentation of variables from intensively sampled areas
- Resampling is done to reduce bias and improve model accuracy



Resampling (cont'd)

- Resampling using the r package spThin (Aiello-Lammens et al., 2015)
- Generates several new data sets that meet a user specified minimum nearest neighbour distance (NND) constraint
- One of the newly generated data sets is used

LAT	LON	SPEC
-17.01	26.69	Uapaca
-11.7963	33.37349	Uapaca
-15.3974	35.39402	Uapaca
-11.4434	34.0367	Uapaca
-14.3333	33.48333	Uapaca
-14.4477	33.87208	Uapaca
-15.3727	35.34496	Uapaca
-15.8115	35.05998	Uapaca

Resampling (cont'd)

- Install the package

```
install.packages("spThin")
```

- Launch the resampling process

```
library('spThin')
```

```
setwd("C:/Users/percy/Documents/R")
```

```
data = read.csv("loc.data.csv")
```

```
thin(loc.data = data, lat.col = "LAT", long.col = "LON", spec.col = "SPEC", thin.par = 100, reps = 10, locs.thinned.list.return = FALSE, write.files = TRUE, max.files = 5, out.dir = "C:/Users/percy/Documents/R", out.base = "thinned_data", write.log.file = TRUE, log.file = "spatial_thin_log.txt", verbose = TRUE)
```

```
thinned <- thin(loc.data = "loc.data.csv", lat.col = "LAT", long.col = "LON", spec.col = "SPEC", thin.par = 1, reps = 10)
```

```
head("loc.data.csv")
```

Formatting output file for MaxEnt



SPEC	LON	LAT
Uapaca	26.69	-17.01
Uapaca	33.37349	-11.7963
Uapaca	33.48333	-14.3333
Uapaca	35.05998	-15.8115
Uapaca	32.125	-19.875
Uapaca	32.875	-19.875
Uapaca	17.96667	-7.21667
Uapaca	29.2	-5.95
Uapaca	39.28262	-8.8308

Species	Y	X
Uapaca kirkiana	26.69	-17.01
Uapaca kirkiana	33.37349	-11.7963
Uapaca kirkiana	33.48333	-14.3333
Uapaca kirkiana	35.05998	-15.8115
Uapaca kirkiana	32.125	-19.875
Uapaca kirkiana	32.875	-19.875
Uapaca kirkiana	17.96667	-7.21667
Uapaca kirkiana	29.2	-5.95
Uapaca kirkiana	39.28262	-8.8308

Correlated variables

- Correlated environmental variables increase collinearity, decrease model transferability, decrease signal to noise ratio, increase computation time and result in inaccurate interpretation of causal relationships
- A pair of variables with a correlation coefficient greater than 0.75 should be considered proxies of each other and one of them should be selected

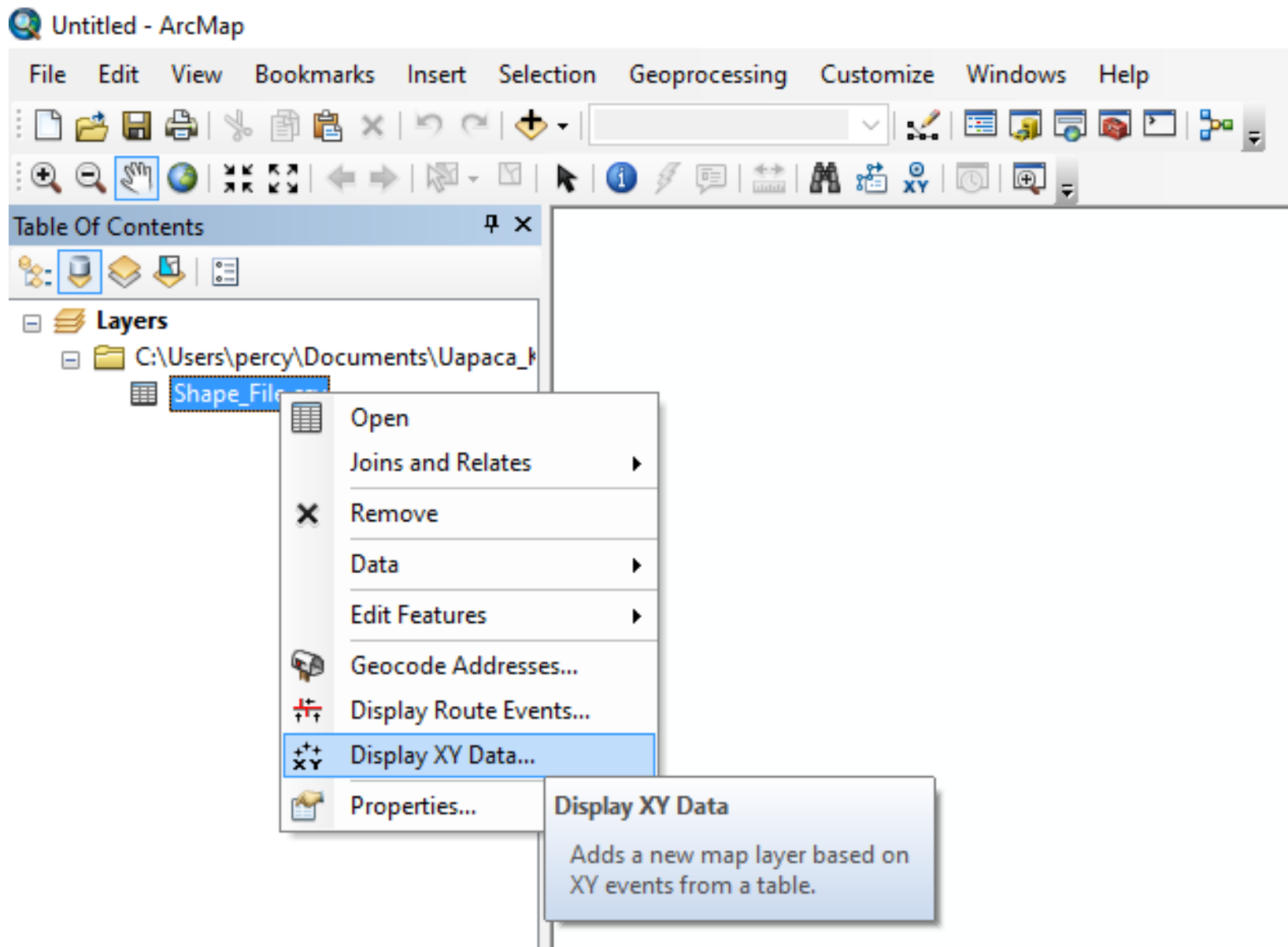
Correlation analysis

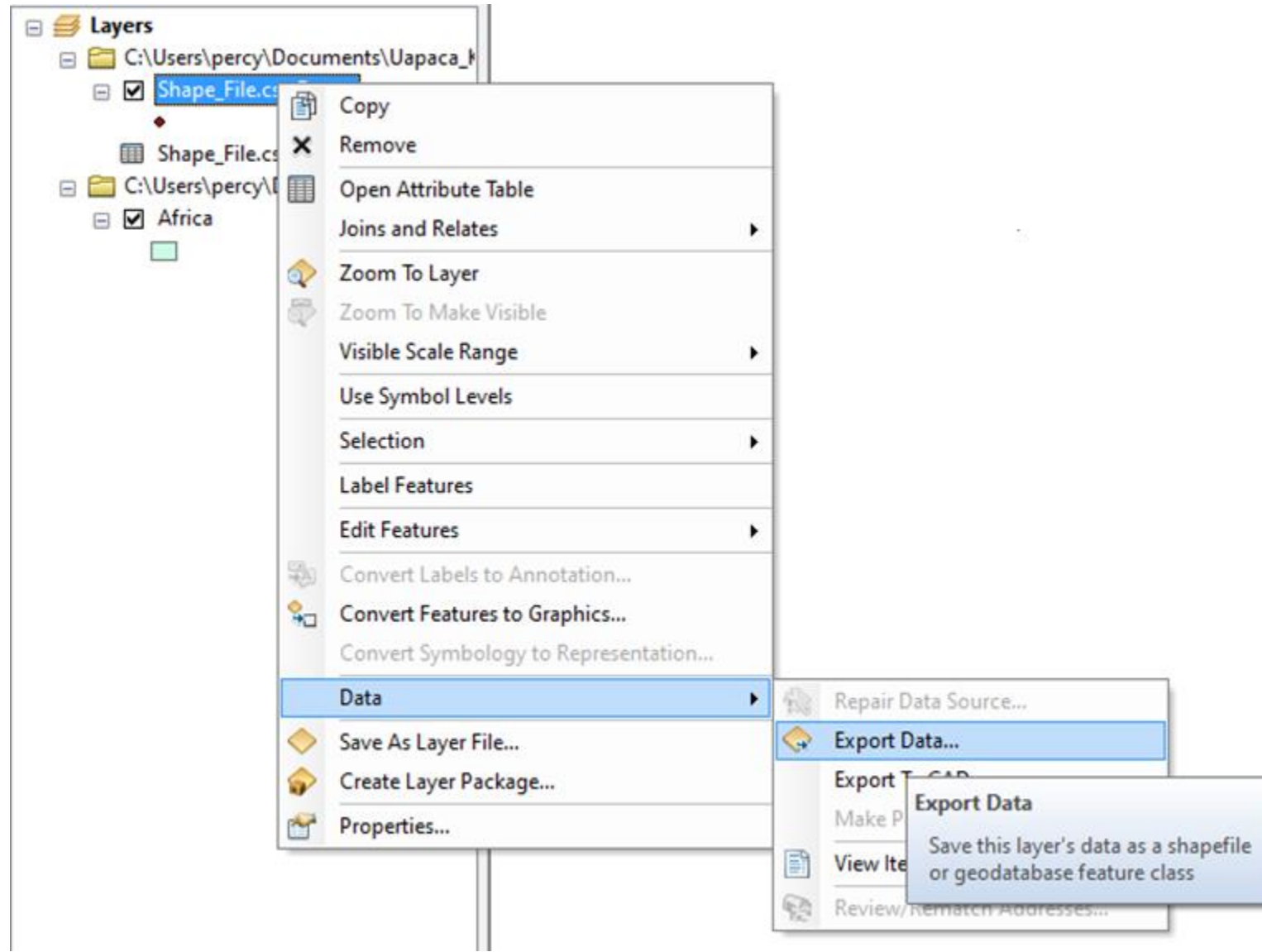
- Correlation analysis in r using the package “caret”
- Prepare output file from resampling as shown and save as csv
- Import the csv into ArcMap/QGIS
- Display the points and save as shapefile

SPEC	LON	LAT
Uapaca	26.69	-17.01
Uapaca	33.37349	-11.7963
Uapaca	33.48333	-14.3333
Uapaca	35.05998	-15.8115
Uapaca	32.125	-19.875
Uapaca	32.875	-19.875
Uapaca	17.96667	-7.21667
Uapaca	29.2	-5.95
Uapaca	39.28262	-8.8308

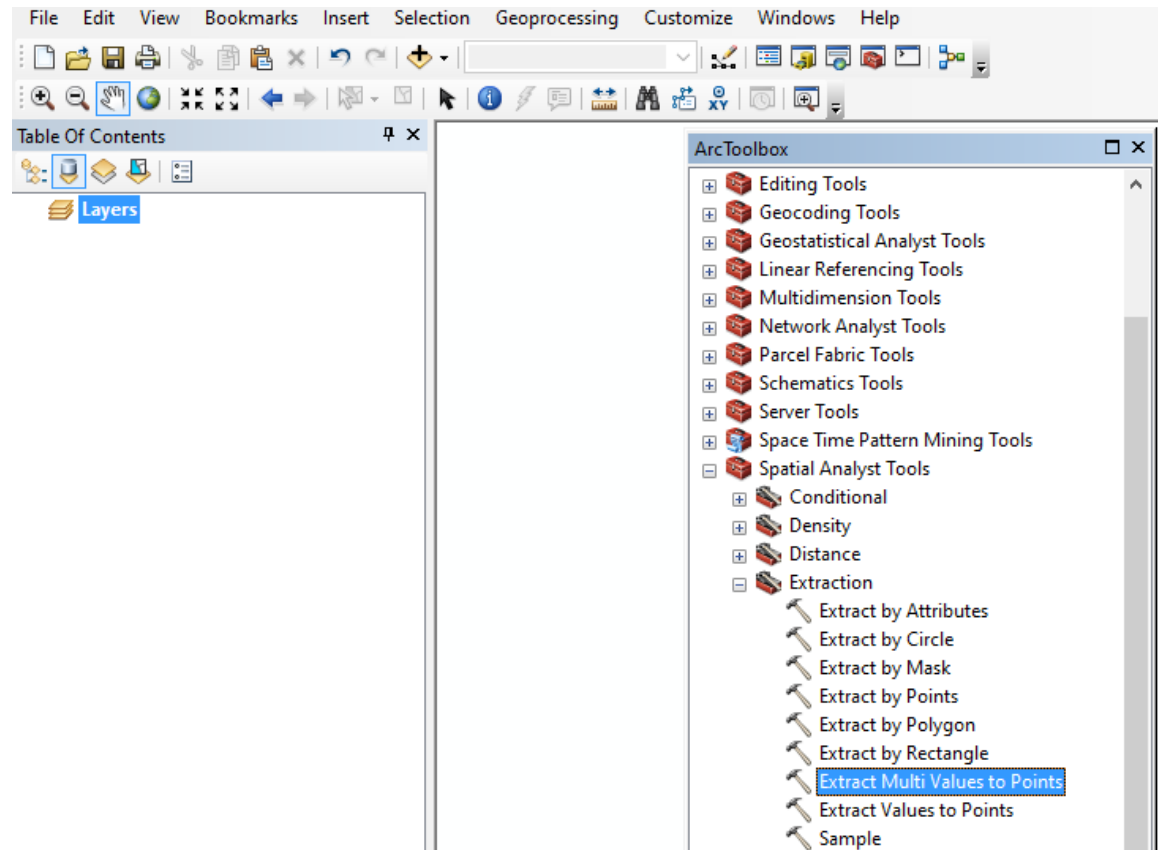


Y	X
-17.01	26.69
-11.7963	33.37349
-14.3333	33.48333
-15.8115	35.05998
-19.875	32.125
-19.875	32.875
-7.21667	17.96667
-5.95	29.2
-8.8308	39.28262





Extract variables from points



Open and format attribute table of shapefile

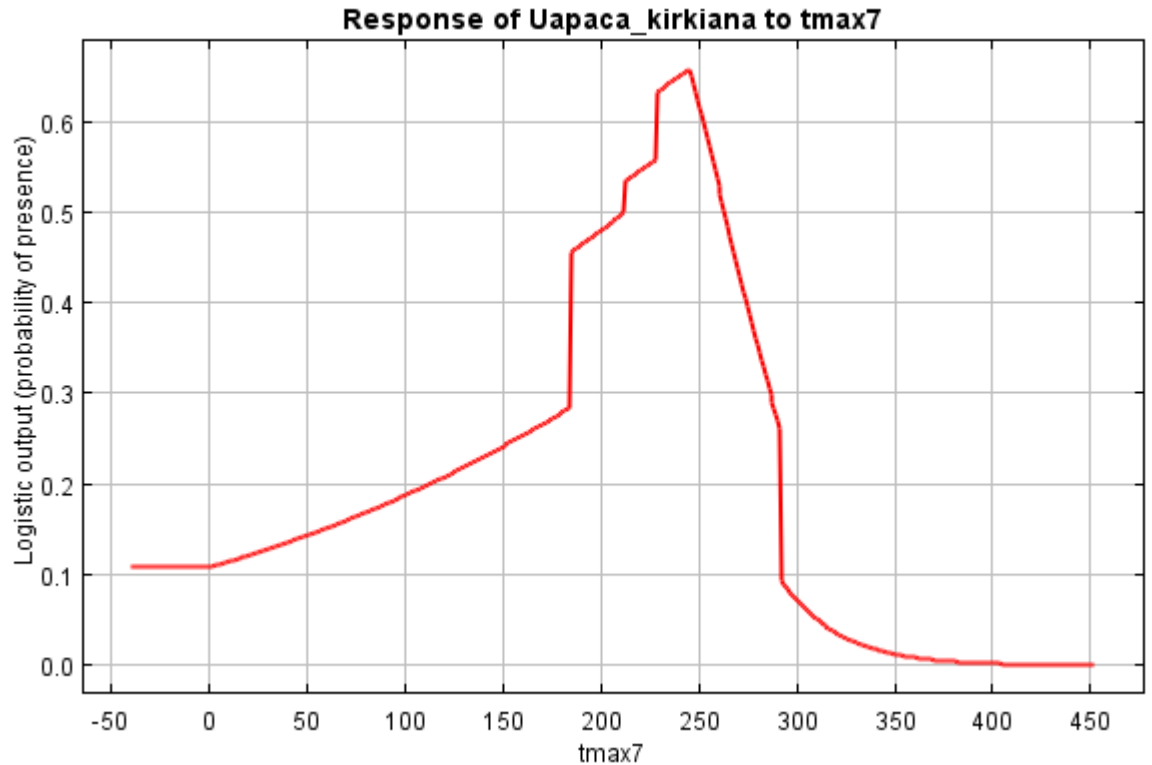
ID	alt	bio1	bio10	bio11	bio12	bio13	bio14	bio15	bio16	bio17	bio18
1	1313	193	226	144	779	194	0	114	550	1	297
2	1160	208	233	175	902	218	0	111	596	3	252
3	1409	191	217	155	1007	238	2	109	664	8	304
4	1164	198	218	168	1187	267	6	97	724	32	367
5	660	214	242	168	477	117	1	101	302	8	282
6	1438	166	189	129	1316	249	22	81	733	74	604
7	868	224	231	212	1542	236	4	65	630	43	565
8	772	236	245	219	1088	206	1	75	459	15	253
9	90	262	273	246	1055	201	8	78	518	31	375

Correlation analysis procedure

```
install.packages("caret") #(Kuhn 2008)
library ('caret')
setwd ("C:/Users/percy/Documents/R")
df1 = read.csv("R_Data_2.csv")
print (df1)
df2 = cor(df1)
hc = findCorrelation(df2, cutoff=0.75) # put any value as a "cutoff"
hc = sort(hc)
reduced_Data = df1[,-c(hc)]
print (reduced_Data)
```

MaxEnt settings

- **Create response curves**- these show how the probability of occurrence of the species changes with increasing values of a variable



MaxEnt settings (cont'd)

- **Jackknife to measure variable importance**- this test shows the contribution of each variable to the model
- The test may be used in conjunction with correlation analysis to reduce the number of variables

Variable	Percent contribution	Permutation importance (%)
Precipitation of the 12 th month	59.6	44.4
Altitude	15.3	19.3
Maximum temperature of the 12 th month	10.4	1.1
Maximum temperature of the 7 th month	5.2	2.1
Temperature seasonality	5.1	8.4
Mean diurnal temperature range	1.9	8.6
Precipitation of the 4 th month	1.4	2.3
Precipitation of the 6 th month	0.8	13.3
Precipitation of wettest month	0.2	0.4

MaxEnt settings (cont'd)

- **Output format**- the default is logistic and it shows the probability of occurrence of a species. The commonly used and easiest to interpret format
- **Random seed**- used when running evaluation runs to randomly partition occurrence between testing and training
- **Random test percentage**- specification of the percentage of occurrence records used for testing and training. Used to evaluate model discrimination ability
- **Replicates**- The number of runs, usually specified when performing evaluation runs

Summary

Obtain occurrence records

Clean and resample

Obtain environmental variables

Correlation analysis

Evaluate the model

Run the model

Analyze output files