HW 5 ~ Species Distribution Modeling

Due Tuesday, Oct. 8th, 2024

Assignment: Like many analyses in spatial ecology, working with SDMs requires that you develop skills to prepare the necessary spatial data sets. In this coding assignment, we will be fitting SDMs to our old friend the Austral grass tree (*Xanthorrhoea australis*). To do so, we will need to assemble two data sets: (1) species occurrence records and (2) environmental rasters. Luckily, you already ahve some practice completing theses tasks. You will first dowload and prepare bioclimatic rasters as in coding assignment #2. You will then download and prepare a *thoroughly cleaned* set of occurrence records for *X. australis*. For these steps, be sure to refer back to your coding assignment #2 submission as well as the solution. After creating a set of background (pseudo-absence) data, will divide the occurrence data into a training set (for model fitting) and a testing set (for model evaluation) and will check our candidate variables for potential issues with collinearity, removing any variables that are problematic before fitting models.

Data 'cleaning' is always important and especially so for data downloaded from online biodiversity databases such as GBIF. Data cleaning typically involves removing duplicates, erroneous records (i.e., records with wrong geographic coordinates, outside of the native range of the species, low spatial precision, etc.). The goal is to produce a data set that is appropriate for (1) fitting SDMs (i.e., to avoid the old modeling adage: 'garbage in, garbage out') and (2) your particular research objective. One important step in the data cleaning process is plotting your data to make sure everything overlaps correctly in geographic space and generally makes sense given what (if anything) you know about the species. Again, refer to the solution to coding assignment #2 to make sure your code works correctly.

You will be graded on your ability to produce clean, well commented R code that performs the tasks listed below without error. When you are done, push your code to GitHub and submit an issue so I know you have completed the assignment.

Let's get started!

Figure 1: A stand of Austral grass trees in Warrumbungle National Park, New South Wales.

---

1. Use the `geodata` package to download bioclimatic variables from the Worldclim (https://www.worldclim.org/) climate data set at 5 arc-minute resolution. Note that unfortunately, we can't use the `worldclim_country` function to download data for only Australia, because, for some reason, that function only provides data at 1km x 1km resolution!

2. In coding assignment #2, we worked with four of the bioclimatic variables. Here, we want to consider more candidate variables for species distribution modeling:

   - Modify the raster stack of the 19 bioclim variables downloaded in step #1 to produce a new stack that contains `bios 2-7`, `10`, `11`, `15`, `18`, and `19` (put another way, the stack should exclude `bio1`, `bio8`, `bio9`, `bio12`, `bio13`, `bio14`, `bio16`, and `bio17`).
   - Crop the resulting raster stack to the *outline of Australia* (not the extent) using the shapefile provided with coding assignment #2. Rename the cropped rasters so they have the correct names (e.g., `bio2`, `bio3`, etc.).

3. Use the `geodata` package to download records for the Austral grass tree (*Xanthorrhoea australis*). Once downloaded from GBIF, clean the resulting data frame by removing records that:

   - Do not have geographic coordinates
   - Fall outside the native range of the species (southeast corner of Australia **only**)
   - Do not overlap the bioclimatic rasters
   - Have coordinate uncertainty greater than 10 km
   - Were collected before 1990
   - Are duplicated
   - Are gridded *spatial* duplicates

As we have discussed in class, gridded spatial duplicates are observations that are close enough in geographic space such that they fall in the same raster grid cell, so spatial duplication depends on the resolution of the raster data. There are some exceptions, but in most cases, we do not want to fit a model using spatial duplicates.

QUESTION 1: In general terms, how would you expect the resolution of a raster to influence the number of spatial duplicates?

At the end of step #3, you should have a cleaned point occurrence data set with the correct CRS. Be sure to plot your data to check if everything seems OK. I ended up with about 500 points post-cleaning. Your result should be close to this number.

After Step #3, you should have (1) a prepared set of bioclimatic rasters and (2) cleaned *presence-only* occurrence records. We need to do a few more things before we are ready to fit and evaluate models. We will be fitting two presence-only SDM methods (Mahalanobis and Maxent). However, our evaluation metrics require absence data and so we will need to generate background points (pseudo-absences) and divide the occurrence data into training and testing sets for the model fitting and evaluation. We also will need to remove highly correlated variables before fitting models.

4. Use your cleaned point occurrence data to extract the bioclimatic variables from the raster stack.

5. Generate a set of 10,000 background points and extract the bioclimatic variables from the raster stack. Combine the resulting tabel with the table produced in Step #4.

6. Use the `vifstep` function in the `usdm` library to remove highly correlated variables from your data table. You will use the remaining variables to fit SDMs. I ended up with 6 uncorrelated bioclimatic variables.

7. Divide the occurrence data into **80% training and 20% testing** partitions using the `ENMeval::get.randomkfold` function.

8. Make a map showing the training and testing presences and background points as different symbols plotted on top of one of the bioclimatic rasters.

9. Use your training data and the uncorrelated set of bioclimatic rasters to fit and predict a Mahalanobis model (using the `mahal` function in `dismo`). Note that `mahal` predictions are slow, so it may take a few minutes to complete the prediction.

10. Make a map of the prediction. The predictions from `mahal` are `1-distance`, so we will need to convert the distance predictions from the `mahal` function to a probability. Here is some R code to do that - it takes as input (1) a raster of the raw prediction (called `mahalPred` in the code below) from the `predict` function and the stack of rasters used to fit the model (called `bioRastsKeep` in the code below).

```
# Convert distances to a p-value
mm.prob <- app(rast(modelPredRast), function(x, k=nlyr(rasterStack)){
  x <- 1-x
  x <- x^2
  p_value <- pchisq(x, df = k, lower.tail = FALSE)
  return(p_value)
})
```

11. Use your training data and the uncorrelated set of bioclimatic rasters to fit and predict a MaxEnt model using `predicts::MaxEnt`. Use jackknifing to assess variable importance and produce plots of variable response curves. Make a map of the prediction. Model fitting could take a moment or two.

QUESTION 2: According to your maxent model, which are the two most important variables associated with the distribution of the Austral grass tree? Which variable is least important?

QUESTION 3: Based on your interpretation of the response curves, what can you say about bioclimatic controls on the distribution of the Austral grass tree?

QUESTION 4: Compare the predicted distributions from the two SDMs. How are they similar / different? Where do the models over- or under-predict the distribution? What might account for these model "errors"?

12. Evaluate the `mahal` and `maxent` models using the testing data.

QUESTION 5: Briefly discuss the model evaluation metrics. Which model performed best? My AUC values for these models were quite similar even though their predictions were not. If you were a conservation manager and were provided output from these two models, how might you handle this seeming contradiction between the differences in the spatial predictions, but similarity in AUC?

QUESTION 6: How might you improve SDMs for the Austral grass tree?