Kylie Blake

Reproducible Workflow Project Proposal

The data I am using comes from my lab, Dr. Jonathan Valente’s Biodiversity Conservation Lab. The previous year, our lab initiated a second iteration of an Inventory and Conservation Planning (ICP) project that the Alabama Department of Conservation and Natural Resources conducted from 2008-2010. 2024 marked the first of a three-year project inventorying vegetation, mammals, and birds on nine wildlife management areas (WMA) across the state. This being my first semester, I am using the vegetation and bird point count data from the 2024 field season to answer an element of my own project’s design. In the previous year, 603-point count sites were inventoried for avian abundance and richness, and vegetation surveys were also conducted to determine landcover type. The dataset is 27,105 entries (one row per bird detection) and 55 total columns.

My project is investigating the role of social information in avian habitat selection for 4 avian species (northern bobwhite, Bachman’s sparrow, prairie warblers, and to be determined). Therefore, I will be using this dataset to inform which WMA has the highest richness and abundance of my target species and the most suitable habitat to install autonomous recording units (ARU) for this social information element of my project.

I will first be using the package *dplyr* to merge the vegetation and bird count data together and then summarize results from the previous 2024 season. I will be calculating species abundance and richness, the proportion of each species per count site, and filtering where target species are by landcover type and WMA. To help visualize the target species’ covers, I will make figures with *ggplot2 a*nd create maps of Alabama to overlay the points on using *sf*, *rnaturalearth*, *dlpyr*, and *devtools*. An alternative, if not too difficult, is making an interactive map, as recommended by my advisor, using *leaflet*. For areas where vegetation data is lacking, I may be creating species distribution models and/ or occupancy models to supplement ARU placement. Once suitable habitat is determined, I will also use *tidyverse, tidyterra, geodata, FedData, tigris,* and *sampling* for point placement. Some possible statistical tests I can use are ANOVA tests to compare environmental variables at controls and treatment sites and permutation tests to ensure site selection is not biased.

To ensure I create a reproducible workflow, I will create a copy of the original ICP bird and vegetation data into my PLPA 6820 GitHub repository as a CSV with a metadata file. I can also input a README file with the directory of the files used in the project and their filepath, as well as any software used required for analysis. For cleaning and merging the data, I will use comments to properly annotate my cleaning and filtering process. For statistical tests and figure-making conducted within R, comments will describe why and how analysis is being conducted, and I will be explicit about what R packages are required. In addition, I will ensure I am pushing all my files listed in the README.txt to GitHub to ensure files are up to date and accessible.