Vedika, Fangming, Alina, Sol

EEB313 Group A Mid-Project Update

Hypothesis:

H0: Water quality change does not have an impact on species richness on birds in marsh habitats.
HA: Water quality change has an impact on species richness on birds in marsh habitats because pollution or other deviations from normal water quality can disrupt normal functioning of organisms.

Predictions:

If we reject the null hypothesis, this means that water quality has an impact on ecosystems. Thus, species richness will decrease as water quality decreases and species richness will increase as water quality increases. If we accept the null hypothesis, then we would predict no differences in species richness among lakes of different water quality.

We have decided to remove amphibians from our research question as we already had over 250 000 rows of data to work with for the birds data. Additionally, we have decided that this dataset is not a reliable way to measure species abundance because of inconsistencies of when observations are recorded. Therefore, we have decided to only compare species richness of birds among sites. Moreover, we have decided to remove the hypothesis on how temperature changes will affect species richness as we had over 1000 water quality traits to compare between. These steps will make our analysis more feasible and allow us to perform more in depth analysis on a smaller number of variables.

Data description, collection, cleaning, and manipulation

NatureCounts is a large biodiversity data repository that contains data on species diversity and abundance gathered by a network of volunteers and scientists. It is the data portal for Birds Canada's National Data Centre, which allows users interested in birds to collect, access and study the datasets on bird species all over the world. Data samples of marshes as well as their bird inhabitants across the geographic area of the Great Lakes Basin, US and Canada were collected since 2004. This data is useful in determining the long-term changes in species abundance and diversity of the bird communities in the wetlands.

The Nature Counts data provides us with dozens of variables that have been measured. However for the purposes of our analysis, we are keeping only a few variables. We are keeping genus and species in order to assess species richness; county, locality, latitude and longitude to compare with water quality data and to understand how species richness changes in different locations; and year, month and day of collection to see how species richness patterns change overtime. We will manipulate the data in R by putting boundaries on latitude and longitude for each lake and using the species that were found in each boundary to measure species richness.

The Government of Canada provides the water quality and surveillance monitoring data from Lake Eerie (2000-2021), Lake Ontario (2001-2021), Lake Huron (2000-2018), and Lake

Superior (2001-2019). Water quality and ecosystem health data was collected in the Great Lakes and priority tributaries to determine baseline water quality status, long term trends and spatial distributions, and the effectiveness of management actions. For each of the 4 lakes, although there are 1004 water quality traits collected, we will clean our data to preliminarily exclude traits with less than 1000 observations, as the total dataset has over 10,000 samples. Some traits that appear to be the most biologically relevant include transmission light profiler (measures turbidity), dissolved oxygen, and pH (Posudin, 2014). However, we will run preliminary analyses to determine which traits may be the most important and explain most of the variation in water quality.

Since each of these traits are provided on a separate row, we will clean up the data by making each quality or nutrient into its own column and the values underneath by converting from long to wide format. We will also get rid of any columns that are unnecessary for our analysis such as ship name, cruise plan, and station number as we will use the lake name and latitude and longitude for the location. Although data past 2018 are not available for all lakes, we will keep all dates as we can compare between just two or three lakes with the data that is available for the recent years. We plan to manipulate the data in R to sort and select the data we need for the analysis such as per year or for a particular nutrient.

Analysis plan and statistical tests/models

Our water quality data is split among several variables that assess water quality, including pH, temperature, dissolved oxygen, and many others. To begin our analysis, we will conduct a Principle Component Analysis to create an axis (or axes) that account for the most variation in We will first run a PCA analysis using the different water quality traits to determine which trait would be the best to assess for further downstream analysis.

To test if there is a difference in species richness among different lakes, we will run an ANOVA. We will also run an ANOVA to test the differences among means of water quality traits among different lakes.

We hope to see the trend of species richness and water quality traits over time before we run any other analyses. To do so, we will run linear regressions assessing the relationship of time and species richness as well as time and water quality. We will also test the relationship of water quality and species richness. Water quality will either be measured using the selected axis/axes from the PCA or by selecting important variables from the results of our PCA that are biologically relevant. We will then run more complex regressions assessing the relationships between multiple variables, which will assess the relationship between both water quality and time with species richness, as well as how the interaction between water quality and time affects species richness.

We will test for all assumptions of each test such as normality, homogeneity of variance, and independent observations. We also hope to also create a map of species richness across lakes to visualize the differences between lakes in our report.

Assumptions:

Despite the availability and usefulness of this Nature Source dataset, we need to consider the reliability of the data as it is a citizen science collection and there may be inconsistencies between observations by different individuals. It is unknown whether all bird observations at a given time and location were reported by the citizen scientist and we are also assuming that all birds in this dataset are marsh birds. We could confirm that the bird species in the dataset are birds that live in marsh habitats by researching some of the species in the dataset and confirming that they are marsh birds. However, this is not feasible to be done for all birds in the dataset, and is an assumption we are taking. Additionally, we are assuming that if the bird is found in a marsh habitat at some point during the year, that they are a marsh bird even though they may be migratory.

Citation:

Birds Canada. 2018. Marsh Monitoring Program. Data accessed from NatureCounts, a node of the Avian Knowledge Network, Birds Canada. Retrieved October 4, 2022, from http://www.naturecounts.ca/.

Government of Canada. N.d. Great Lakes Water Quality Monitoring and Surveillance Data.. Retrieved October 4, 2022, from https://data.ec.gc.ca/data/substances/monitor/great-lakes-water-quality-monitoring-and-aquatic-ecosystem-health-data/great-lakes-water-quality-monitoring-and-surveillance-data/

Posudin, Y. 2014. Measurement of Water Quality Parameters. In Methods of Measuring Environmental Parameters, Y. Posudin (Ed.). https://doi.org/10.1002/9781118914236.ch20