DATA C200 Project Report:
Biodiversity in National Parks

April 29, 2022

Alora Clark

Please Note: this project is starkly different from the project originally planned and proposed in the feedback form. *

ABSTRACT

In the United States, national parks are places designated for the conservation of wildlife, both plants and animals. The biodiversity of these areas can be affected by many environmental factors, natural and man-made. In this project, I am to answer the question: "Can we predict the biodiversity (total number of present and approved unique species) of a national park based on varying common attributes of a national park (such as location, size, etc.)?" To assess the answer to this question, I make use of exploratory data analysis and linear regression modeling to make predictions. The answer to this question is important because it can potentially further inform decisions regarding national parks such as construction, removal of a species, invasive species, expansion, etc.

1. INTRODUCTION

This project is concerned with the biodiversity found within a national park. Biodiversity is defined as the variety of life in the world or in a particular habitat or ecosystem. For the context of this project, I am referring to biodiversity as the total number of unique present and approved species located in a given national park.

Because the ability of a species to grow and/or thrive is often influenced by environmental factors, I want to discover the answer to the question: "Can we predict the biodiversity (total number of present and approved unique species) of a national park based on varying common attributes of a national park (such as location, size, etc.)?"

The answer to this question is important because national parks are places of preservation. By knowing what factors potentially affect biodiversity, we can properly account for and predict how biodiversity will be impacted based on changes to the park such as construction, expansion, removal, the introduction of non-native species, etc.

Many people use biodiversity to predict the likelihood of a species becoming endangered or when the species will become endangered. They often use what is called species distribution modeling (SDM). Species distribution modeling is “predict[ing] species distributions by identifying key environmental (i.e., abiotic) characteristics of suitable species habitats and then [using] models that incorporate both information on known occurrences of a species (i.e., presence data) along with certain environmental characteristics of those known habitats in which the species occurs to estimate potential habitats in locations where species occurrence data are lacking.” [1].

From this information and prior studies, we can assume that it is within scope to use environmental factors to predict biodiversity. This research differs from others, because we are not predicting potential habitats, but rather we are predicting biodiversity/species richness which the number of different species represented in an ecological community, landscape, or region. Here, we are not accounting for the various distributions of said species.

I pursued the answer to my main research question by first doing some exploratory data analysis to select my features and then I created the model. The methodology for this project will be further expounded upon in section 2.

2. DESCRIPTION OF DATA

This data was collected by the National Parks Service in the United States. Their purpose of creating this dataset was to "provide information on the presence and status of species in our national parks."

According to the National Parks Service's description of this data, the "time and effort spent on species inventories varies from park to park." This can mean that just because an animal is not seen within the park, that does not mean it does not reside there. Likewise, if an animal is seen within the park, it could be a rare occurance and the case that the animal does not reside there. This could potentially lead to a sampling bias known as undercoverage, where some members of the population are inadequately represented in the sample.

For this project, I chose to use the ‘national_parks_biodiversity_parks.csv’ file and the ‘national_parks_biodiversity_species.csv’ file. The ‘national_parks_biodiversity_parks.csv’ file contains data about 55 national parks within the United States. The name of the park, the park code, the state the park resides in, the location (longitude and latitude), and the size of the park (in acres) are all included in this file. Additionally, the ‘national_parks_biodiversity_species.csv’ file contains all of the species information for each species in each of the national parks. This data also has a high level of granularity as the species-id, category, order, family, scientific name, colloquial name, and several other variables are all included in the table.

2.1 Exploratory Data Analysis

At first glance of the data, I noticed that the ‘national_parks_biodiversity_species.csv’ file had several columns with a significant amount of missing data. Because these columns were not pertinent to my research question or analysis I decided to drop them from the overall table. I also noticed that the ‘Occurrence’ and ‘Record Status’ columns contained several categorical labels that could be interesting. ‘Occurrence’ contained the labels ‘present’, ‘not present’, ‘not present (historical report)’, ‘not present (false report)’, ‘in review’, and ‘not confirmed’. Because we have defined biodiversity to be the current total number of unique species, we only want to consider the species associated with the ‘present’ label for each park. Subsequently, ‘Record Status’ contains the labels ‘approved’, ‘in review’, and several others and for the same reason specified above, we only want to consider those species that have been approved. Lastly, there were a few NaN values in some of the columns, but I chose to deal with these after the overall transformation of data.

After cleaning up the data, I created one single data frame with all of the columns from both files. This data frame allowed for easier splitting and creation of sub-data frames. To answer our research question, we must first answer several sub-questions and identify if there is a correlation between the total number of unique species and the variables of interest. The sub-questions I considered are as follows:

1.) Is the number of different species based on the location of the national park (i.e. is there a correlation between the unique number of species and the longitude/latitude of each park)?
2.) Is the number of different species correlated with the number of acres that the park occupies?

The interest in sub-questions 1 & 2 is easily understood as they directly concern the relationship between the total number of unique species in a park and our variables of interest. The answer to this question will also help to inform us of whether or not these variables, location and size, will make good features for our model.

2.2 Interesting Findings

To answer our first two research sub-questions, we needed to determine if there was a correlation between the variables. In order to do this, I had to add a column to the data frame that I created from ‘national_parks_biodiversity_parks.csv’. This column contained the total number of unique species found in each of the national parks. Once I did this, I plotted this number against the size of the park to create a scatter plot, where each point represents a park.

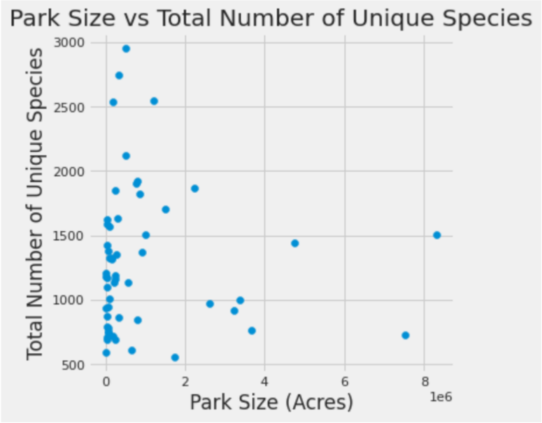

Figure 1. Scatter plot of park size vs. the total number of unique species

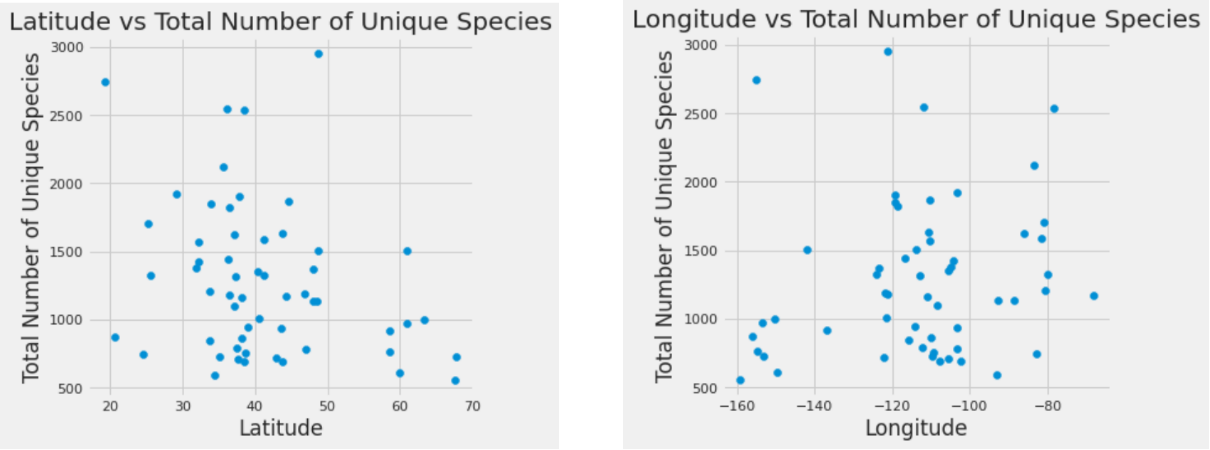

Figure 2. Scatter plots of longitude and latitude vs the total number of unique species

As we can see from figure 1, the scatter plot is skewed to the right and seems to be overcrowded. This makes it very difficult to see if there is a correlation between variables, and even if we could successfully visualize the correlation, we do not know what the strength of the relationship is. To find the strength of the relationship, I decided to use a Spearman's r test. This test is like Pearson's r but does not require prior knowledge of the distribution of variables (i.e. you do not need to know if the variables are normalized or follow a normal distribution). Our Spearman's r coefficient was **0.24** and indicates a weak positive linear relationship between the variables.

This was extremely interesting to me as I thought that size would weigh heavily on the biodiversity of a national park. Another interesting observation from this plot is that it appears the larger a park the less biodiversity it has, based on how we have defined biodiversity. This could be due to the fact that this data set might suffer from undercoverage (note: this is briefly mentioned in the summary of the description of the data above) and because the park is so large, researchers were not able to locate all of the unique species that inhabit the park.

I also plotted the total unique species values for each park against the longitude of the park and the latitude of the park. The scatter plots gave a clear indication of a positive correlation and negative correlation, respectively. However, we could not discern from the visualization what the strength of the relationship is. I conducted a
Spearman's r test for this data as well and the results for the Spearman's r coefficient were **0.138 and -0.274**.

These values indicate a weak correlation between variables which was also surprising to me. This is because I initially assumed that location would have a strong correlation to biodiversity due to the fact that different locations offer different climates, which may be better for some species than others.

3. DESCRIPTION OF METHODS

3.1 Methods Used

To answer my research question: "Can we predict the biodiversity (total number of present and approved unique species) of a national park based on varying common attributes of a national park (such as location, size, etc.)?" I first needed to know if any of the variables I was considering were correlated with each other and/or biodiversity. Because of this I completed EDA (exploratory data analysis) on several of the variables and created many of the above plots for biodiversity vs size and location. This gave me a good idea of what relationship each of these variables had to biodiversity and if they would be good features for my model.

The next part of the process was concerned with actually building the model and running predictions. This is covered in more detail in section 3.3.

3.2 Causal Inference

Although important, this project does not rely heavily on causal inference. Causal inference is defined as the process where causes are inferred from data. [2] Compared to correlation, where we observe the probability of x given y, in causation, we observe the probability of x given that y was done. [2] Additionally, correlation allows us to understand if there is a relationship between variables, whereas causation allows us to understand which variable is causing which. The drawback of causal inference is that there can be many uncontrolled variables that influence the outcome, also known as confounders.

3.3 Modeling

I decided to build a multiple linear regression model because we are trying to predict a quantitative value, the total number of unique species, using more than one feature. I then selected my features based on the EDA conducted earlier and I decided to include the additional variables of ‘Park Name’, ‘State’, ‘Occurrence’, and ‘Category’. Because all of these variables are categorical variables I had to do one hot encoding for each of them and I dropped the first column of each of the variables (that were one-hot encoded). I noticed that including all of the category values as columns caused overcrowding and inaccuracy in the model so I only decided to include ‘Fungi’, ‘Mammal’, ‘Nonvascular Plant’, ‘Reptile’, and ‘Spider/Scorpion’.

I then split my data into a training set and a test set. Next, I created a multiple linear regression model using sci-kit learn, fit the model to my training data, and used the model to make predictions on my testing set. I then computed the test error for my model using Root Mean Squared Error and adjusted-R^2 to assess the performance of my model. Finally, I plotted the predicted versus the observational values and the predicted vs residual values as another assessment of the appropriateness of the model.

4. SUMMARY OF RESULTS

4.1 Inference

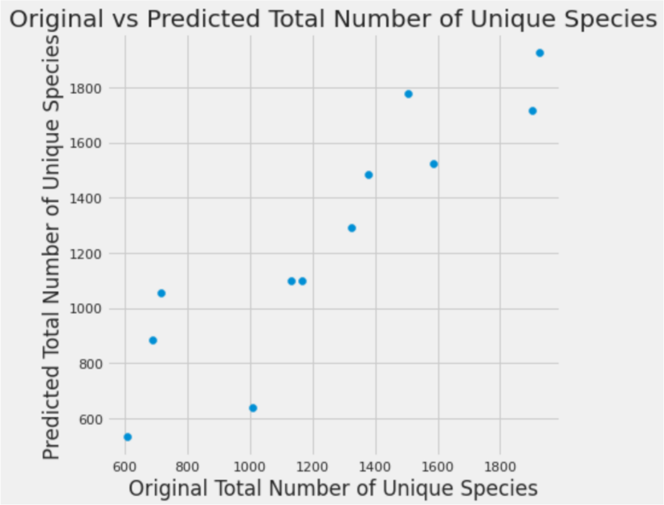

Figure 3. Observed vs Predicted values for the Total Number of Unique Species in a national park

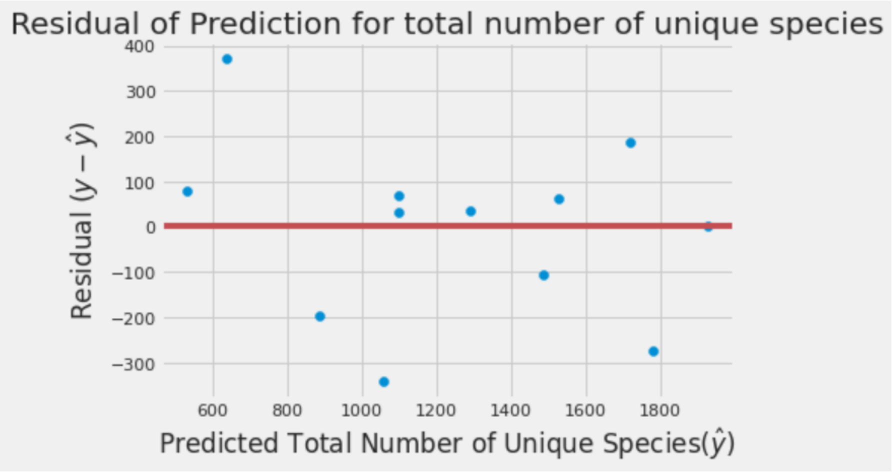

Figure 4. Residual Plot

The multiple linear regression model that we have created seems to be a good model for predicting the biodiversity of a national park. We can assume this is a good model for our predictor because our adjusted-R squared is 0.802 and our RMSE is approximately 190, which given the range of values for our predictor is decently low. We can also tell that this model is a good model for our predictor because of the original vs predicted values plot (figure 3) and the residual plot (figure 4). Observing the original vs predicted values plot we can see that the data follows a positive linear trend and is moderately close to the y=x line. Additionally, our residuals plot has few outliers, and most of the data clusters around +/- 100, which again is pretty good given the range of values for our predictor.

4.2 Comparison of Models

Before I settled on a Multiple Linear Regression Model, I tried using an Ordinary Least Squares Linear Regression model. This model is best for estimating the coefficients of linear regression equations, which help to describe the relationship between one or more quantitative independent variables and a dependent variable. To complete the OLS regression I used the statsmodels.api library in python, and when I initially did the OLS regression I only had 3 features – longitude, latitude, and size of the park to predict the biodiversity. These features worked fine to produce a result, but this type of model did not work well for making predictions and I knew I needed to add more features to the model to make accurate predictions of any kind.

I then attempted to do OLS regression again after creating my multiple linear regression model with the features utilized in the multiple linear regression model. This yielded an extremely unfavorable result as many of the statistical values were NaN or undefined. I believe this was due to the fact that OLS works best with quantitative variables and many of my features were categorical variables.

4.3 Limitations

Overall, I think using multiple linear regression was the right approach to answer my research question. This model allowed me to make predictions about the biodiversity of national parks and select the best features to do so. The main limitation with this approach, and really any linear regression model, is assuming linearity between the dependent (predictor) and independent (feature) variables. Rarely, if ever, is data linearly separable, as there is usually not a straight-line relationship between variables. Additionally, I believe that this model could've benefitted from more quantitative data such as temperature, the number of visitors in any given time period, average rainfall, etc. The categorical variables work well to help the accuracy of the model, but these additional variables could've informed the relationship between the environment of a national park and the biodiversity of the park a bit better.

4.4 Future Work

I was surprised to discover that 'Park Name' would be a good feature for my multiple linear regression model. When I trained the model without the 'Park Name' feature, the RMSE was very high and the adjusted-R squared was extremely low. This result was surprising, because the name of the park is not technically an attribute of the environment for that park and I don't believe is highly correlated, if at all, with the total number of unique species within the park. I suppose this is something that I can expand upon in the future, to discover why this feature was a good feature for my model.

I was also surprised that 'Mammal' was a good category to have as a feature and something more obscure like 'Crab/Shrimp/Lobster' was not. This is surprising because all of the national parks are very likely to have some number of mammal species, as that category is very broad and encompasses many organisms. However, something like 'Crab/Shrimp/Lobster' is less likely to be found in all or most of the parks, which I thought would make it a good distinguishing feature when predicting the number of total unique species within the a park.

As stated in section 4.3, I think that in the future this model could benefit from weather data (i.e. temperature, rainfall, snowfall, etc.) in order to better predict the biodiversity of a national park. Similarly, it might be helpful to know which seasons (winter, spring, summer, or fall) the park experiences and which biomes (dessert, wetlands, etc.) the park encompasses to better predict the biodiversity. Lastly, I think this dataset could benefit from the collection of the biodiversity of each park over a period of time (for example year to year). This would be a great feature for the model and could increase its accuracy greatly.

Future research could include the collection of the above variables, as well as expanding upon the research question to possibly try and predict the number of invasive species or the number of endangered species within a national park.

5. CONCLUSION

We have considered the research question: "Do the attributes of a national park, such as the location, size, etc. influence the level of biodiversity within that park?" and "Can we use these features to predict the biodiversity of national parks?" Based on the above analysis we can say that there is some correlation between the size of a national park, the location of the park, and its biodiversity. However, this correlation is not strong and should be evaluated as such. Based on the evidence above, we can conclude that the attributes that make up a national park do contribute to some extent to the biodiversity of that park, but further data and analysis is needed to truly understand what the causal relationship is between these variables and the biodiversity of national parks. Furthermore, we can conclude that we can use the features of a national park to predict its biodiversity. We can draw this conclusion, because we have created a multiple linear regression model that does exactly this, and we know that it is indeed a satisfactory model because our adjusted-R squared is greater than 0.75-0.8 and our RMSE is considerably low given our range of total unique species in our dataset.

REFERENCES

[1] “Species Distribution Modeling Data.” *NASA*, NASA, 7 Apr. 2020, https://earthdata.nasa.gov/learn/pathfinders/biodiversity/species-distribution.

[2] Wang, Kevin. “Implementing Causal Inference: Trying to Understand the Question of Why.” *Medium*, Towards Data Science, 30 Mar. 2020, https://towardsdatascience.com/implementing-causal-inference-a-key-step-towards-agi-de2cde8ea599.