Analyzing the effects of Canopy Height and Species Richness through LiDAR Data

Nikki Shukla, Jackson Stachnik, Jordan Sheehan, Lindsay Goldberg-Custer, Sean Davis

University of Florida

IDS2935: Can Big Data Save the Earth?

Geraldine Klarenberg

April 27, 2022

Introduction

A significant component in environmental science studies is determining the relationships between the different plant species of an ecosystem. For example, forests that are over 200 years old contain taller and larger trees. Therefore, they cover a larger amount of the terrain. Eventually, if these forests are not well-maintained, their trees will overgrow. This will reduce the amount of sunlight that reaches the plants underneath the canopy. Consequently, researchers are making efforts to minimize this problem with controlled burns and thinning. Overall, our research can help us to better understand which types of ecosystems and their trees can best help the smaller plants underneath flourish. The primary variables that we will analyze include tree height and species richness. We will also be looking at climate conditions, including average precipitation, wind speed, humidity, and temperature.

Establishing a model for the variables that affect species richness in a given ecosystem is complex; however, through the use of statistical coding packages we are able to draw individual correlations between all the variables that we chose to describe. In this study, we will be observing three different NEON research sites: Harvard Forest (HARV), Niwot Ridge (NIWO), and Ordway Swisher (OSBS). We researched the extent to which tree height correlated with plant species richness. We hypothesized that there will be a moderately negative correlation between tree height and plant species richness because greater tree height means there will be less light going through the canopy for photosynthesis to support species richness. This data is collected through various means, but our main aerial data was collected using LiDAR (Light Detection and Ranging).

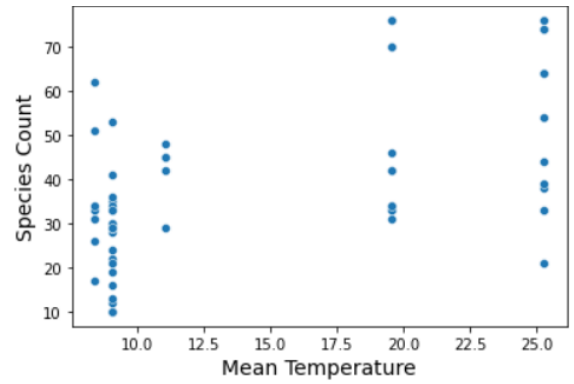
Methods

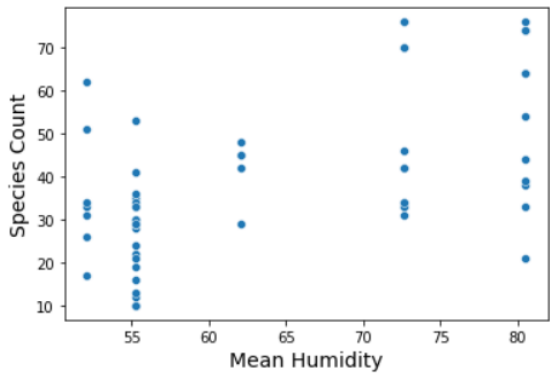
The first step in our analysis was to select data for our research question. We chose three different research sites (HARV, NIWO, and OSBS) because we felt we could produce the best results by analyzing data from a variety of ecosystems. We then chose several weather variables which we thought may have an effect on species richness and included them into our main data set. For our code, we imported the NumPy, Pandas, Matplotlib Pyplot, Seaborn, and SciPy packages. We then loaded in all the data into dataframes separated by individual variable (weather variables, species richness, tree height) from their CSV format. Before moving forward, we spent time cleaning the data so it could be better used for future analysis. This involved removing all observations that were flagged for poor quality and creating a properly formatted date column. Next, since NEON only collects species richness and tree height data at specific days during a site's greenest season and weather data is collected every day, we had to create a workaround for combining the two dislike dataframes. We accomplished this by calculating the mean for each weather variable by site in the respective months that species richness and tree height data were also collected in. We then added these mean values by matching them with the appropriate site in the species richness dataframe. After this was completed, we moved on to wrangling our species richness data. We counted all unique species, genus, family, and variety

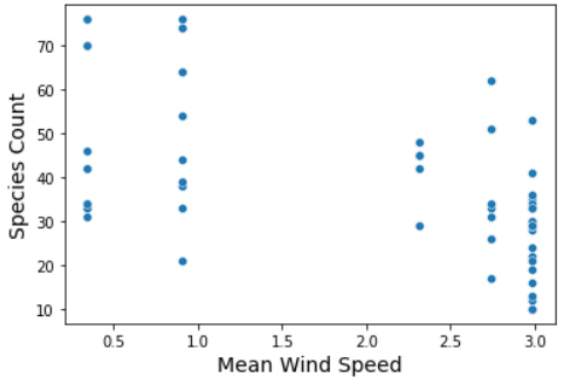
identifiers per plot to serve as our species richness indicator. We called this variable "speciesCount" and added it to the dataframe. The last step of manipulating the data involved merging the tree height data and further cleaning the final data set for use in our analysis.

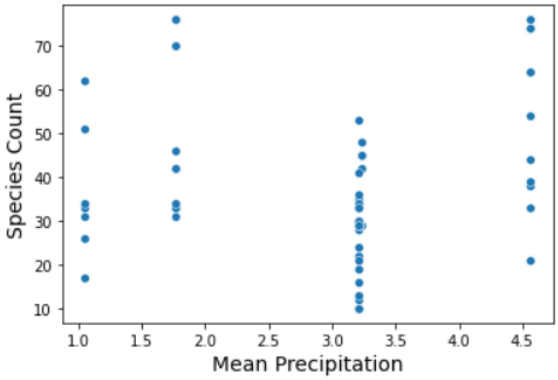
Analysis and Results

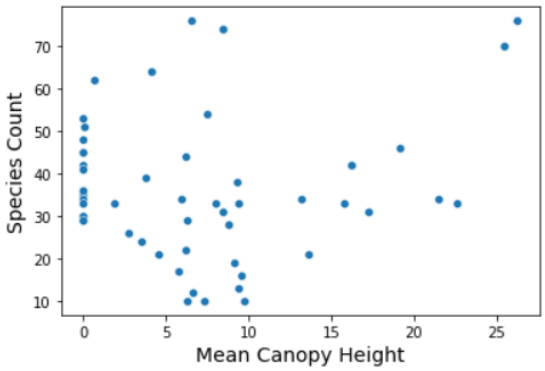
To begin our analysis, we decided to create scatter plots using the Seaborn package to get a surface-level understanding of the relationship between each of our independent variables and the dependent variable (species count). Through plotting each weather variable (mean humidity, temperature, wind speed, elevation) and tree canopy height against the species count, we created the following scatter plots:



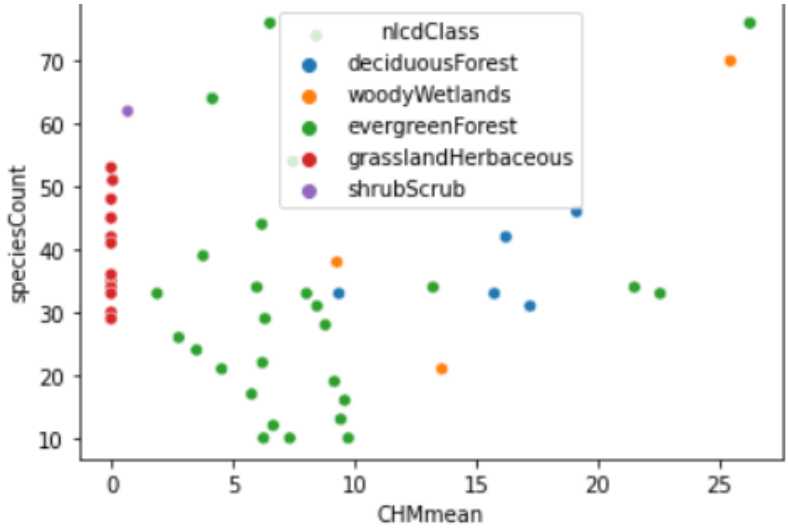




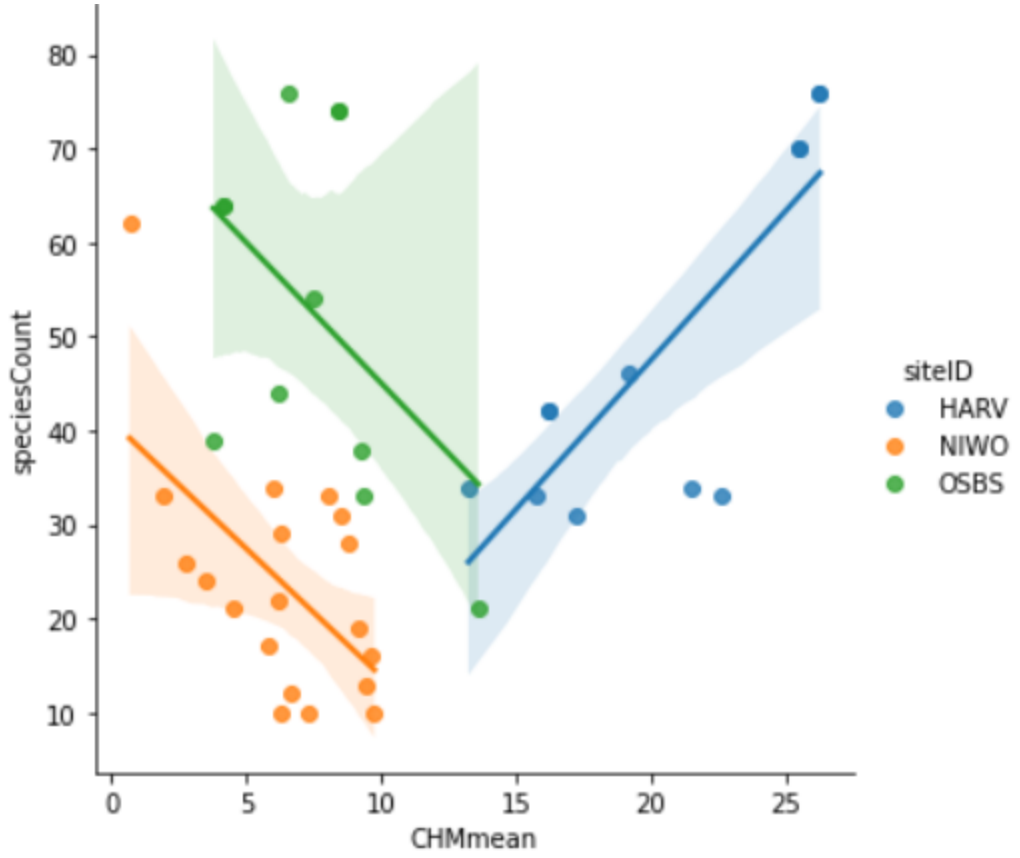




Evidently, we are not able to draw any definitive conclusions about species count from these visualizations because the data points were scattered and seem randomly placed. We concluded that there is no evident relationship between any weather variable and species richness, as the species count varied greatly even across similar weather conditions.. After ruling out the weather variables as strong factors, we shifted our focus to the canopy height data. To help with our visual analysis, we decided to separate this data by biome type. By assigning the "nlcdClass" variable (which indicates the vegetation type of the plants) as the hue, we generated this plot:



Because herbaceous grasslands had minimal tree coverage, we decided to remove that data before computing linear regressions comparing canopy height and species count. Also, as we continued our analysis, we looked at data from each site separately because all three sites combined contained lots of evergreen forests, which prevented us from analyzing areas with less data such as deciduous forests and woody wetlands. Pictured below is the linear regression plot:



This plot assigns hues based on the site our data was sampled from, and gives a linear regression for each subgroup. The shading represents the statistical error of our linear regression. Though at an initial glance there seems to be correlation in each of these subgroups, the high error we computed– especially for the OSBS data– muddles our results. Our plot also indicates a negative correlation for the NIWO and OSBS sites but a positive correlation for the HARV site, but these findings cannot be as indicative of a pattern due to the error. Therefore, we couldn't gather support for our hypothesis regarding there being a negative correlation between canopy height and species richness or make any other accurate conclusion due to the error. It is possible that other geographic factors may have influenced species richness more than canopy height. Looking back at all of the canopy height data, we later computed the R-squared value of the linear regression to be 0.11. This value indicates that there is a weak association between species richness and canopy height. In conclusion, we reject our hypothesis.

Word Count: 982