# **Plants in the Human-Altered Environment (PHAE) Research Project**

Module 7: Data Exploration

Now that you have entered your data into the standardized format for the national research project developed by EREN-NEON, we are ready to begin analysis of some of the data by comparing woody plant species richness and diversity across our plots. We’ll also examine whether there is a relationship between the degree of human disturbance in a plot and the woody plant diversity in that site.

In this module, we will focus on our class data, so the climate data is not particularly relevant as most of our sites are located in roughly the same area. The climate data will be more important for future analysis using the broader NEON dataset, and will be valuable for future studies that build on the work you started this semester.

Part 1: Calculate site summary statistics

Our class sites are evenly divided into half that are designated by the NLCD land cover classes as developed to some degree (those that begin with the word “developed” in column H on the provided data sheet) or more natural (those that don’t), an ideal distribution for some broad comparisons!

Use the provided google sheet specifically for this module (it is condensed for simplicity). Begin by calculating some basic summary statistics to complete the table below.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Developed Plots** | **Not Developed Plots** |
| **Average** impervious surface |  |  |
| **Average** semipervious surface |  |  |
| **Average** lawn surface |  |  |
| **Average** tree cover |  |  |
| **Average** % measured impervious surface (calculate % using column D/column C, then take average) |  |  |
| **Average** % measured tree cover (calculate % using column G/column C, then take average) |  |  |
| **Average** NLCD Impervious designation |  |  |
| **Average** NLCD Tree cover designation |  |  |
| **Average** VH humanmade % |  |  |

**Question: How do the NLCD designations compare with the measured estimates of land cover? By taking the average among sites, do the NLCD values get more accurate (closer to the true value) or less as compared to each site individually? Why?**

Part 2: Calculate tree summary statistics

Begin by calculating the species richness and total number of trees in each plot (second page on the provided data sheet), then find the average for each site type to fill in the table below. Use the STDEVP function in the google sheet or Excel to get the standard deviation.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Developed Plots** | **Not Developed Plots** |
| **Average** number of trees per plot |  |  |
| **Average** species richness per plot |  |  |
| **Average** tree diameter (all trees) |  |  |
| **Standard deviation** number of trees per plot |  |  |
| **Standard deviation** species richness per plot |  |  |
| **Standard deviation** tree diameter (all trees) |  |  |

Part 3: Create bar graphs

The averages and standard deviations in our dataset provide information about the power and relevance of the data. A bar graph is a simple way of visualizing the data to allow for comparisons.

The average of values in a sample (like the average number of trees per plot in a developed area) is essentially a general statement about what we would expect to find in a new plot like the ones that we sampled. The standard deviation helps us to understand how good this estimate would be – is the average really representative, or is there a lot of variability? In a bar graph, the standard deviation is used to create error bars, a visual representation of the measure of variability.

1. If you have not already, copy the table from Part 2 into Excel (you can also use a new google sheet, but the instructions will be slightly different). It may be helpful to create three separate tables, one for each variable, like this example:

|  |  |  |
| --- | --- | --- |
|  | Developed plots | Not developed plots |
| Average number of trees per plot |  |  |
| SD number of trees per plot |  |  |

1. Create bar graphs for each of the three variables. Select the cells with the average values 🡪 click Insert 🡪 click the 2-D bar graph icon in the middle of the top of the page. Click “Add Chart Element” in the top left and click through the options to adjust your graph with appropriate title, labels, axis titles, etc.
2. Add error bars. Click “Add Chart Element” again 🡪 select “Error Bars” 🡪 “More Error Bars Options” 🡪 under “Error Amount” select “Custom”. Click “Specify Amount” 🡪 highlight the cells containing your standard deviations 🡪 click OK.
3. Double check your bar graphs. Each one (three total) should:

* Have two bars (one for developed, and one for not developed)
* Have the x-axis labelled with the type of plot (developed and not developed)
* Have a y-axis label indicating the variable
* NOT have gridlines (click on one of them to delete, they are unnecessary clutter)
* Have error bars that indicate the standard deviation that you calculated above (NOT Excel’s automatic error bars which are incorrect – double check that the size of the bar reflects your value).

**Copy and paste your completed bar graphs below.**

Part 4: Make hypotheses and create scatter plots

Now that you have done some initial data analysis on our PHAE plots, we can start testing some hypotheses. You measured two key parameters of human impact in your plots, the “viewshed” impact (% humanmade visible) and the impervious surface impact.

Make some hypotheses about the relationships between parameters you measured across ALL plots (not separated by developed or not developed). You can use any of the data summarized in parts 1 and/or 2. For example, one common hypothesis was “the total number of trees will decrease with increasing impervious surface.” Choose two different pairs of variables to compare across sites.

**Write your two hypotheses here:**

1. Create two tables in Excel (or google sheets, but the instructions will be somewhat different) by copying the relevant data from the provided google sheet. For example:

|  |  |  |
| --- | --- | --- |
|  | # trees per plot | % impervious surface |
| Mead Gardens 1 |  |  |
| Shady Park |  |  |
| Knowles Garden |  |  |
| etc. |  |  |

1. Select the cells containing the data. Click Insert 🡪 click the scatter plot icon (middle of the bottom of the options where bar graph was the top left in Excel). Double check that your independent variable is on the x-axis and dependent variable is on the y-axis; if it isn’t, click “Select Data” at the top of the page and switch it. Click “Add Chart Element” to edit titles, axes, etc. as before.
2. Add a trendline. Click “Add Chart Element” 🡪 Trendline 🡪 Linear.
3. Evaluate the relationship between your variables. Click “Add Chart Element” 🡪 Trendline 🡪 More Trendline Options. Check the box next to “Display R-squared value on chart” at the bottom.

**Copy and paste your two scatterplots here**

**Question: The R-squared value is a measure of how strong the relationship is between two variables. It is a number between 0 (there is no relationship between the variables) and 1 (there is a 100% correlation between the two variables, a very strong relationship). What do your R-squared values tell you about your hypotheses? Does this data support your hypotheses? Why or why not?**