**Module Adaptation Overview**

Building research questions and carrying out appropriate analysis to answer them is often the ultimate semester activity within introductory ecology courses. These research experiences present multiple challenges concerning 1) how to generate data and 2) how to manage the data for analysis. This adaptation uses NEON vegetation structure data in place of NEON small mammal data in the teaching module '*More in Depth Spreadsheet Management Adaptation of Data Management Using NEON Small Mammal Data'* (Raisa Hernández-Pacheco 2018)

 to address the second challenge.

This activity is designed to be carried out in a 1.25 hr lecture style course, however, students will need access to computers during this activity. Although the activity can be easily extended, its adaptation to the original was intended for instructors with high need to teach data management skills in a short period of time.

**Specific module adaptations**

NEON terrestrial forest vegetation data replaced NEON small mammal data in order to broaden the use of the teaching module in plant oriented courses. Specifically, this adaptation consists of four exercises that introduce students to 1) format spreadsheet data tables, 2) carry out spreadsheet quality control, 3) count/sort/filter data of interest in order to conduct a pilot analysis on NEON forest aboveground biomass, and 4) correlate vegetation biomass across nine forest sites to a climate metric (e.g., latitude, mean annual temperature, or mean annual precipitation).

In this adapted activity, students will learn how to:

* format spreadsheets for effective data use.
* carry out spreadsheet quality assurance and control.
* conduct a pilot analysis by managing a large, open access dataset to answer a research question.
* create a scatter plot figure correlating tree aboveground biomass to a climate metric across nine sites.

**Original modules**

This lesson is an adaptation based on another teaching module:

Hernández-Pacheco, R. (2018). [More In Depth Spreadsheet Management Adaptation of Data Management using NEON Small Mammal Data](http://dx.doi.org/10.25334/Q44X4D). [NEON Faculty Mentoring Network](https://qubeshub.org/groups/neon2018), QUBES Educational Resources. [doi:10.25334/Q44X4D](http://dx.doi.org/10.25334/Q44X4D)

The Hernández-Pacheco teaching module is also an adaptation based on two other educational modules:

*Christie Bahlai and Tracy Teal (eds): “Data Carpentry: Data Organization in Spreadsheets*

*Ecology Lesson.” Version 2017.04.0, April 2017,* [*http://www.datacarpentry.org/spreadsheet-ecology-lesson/*](http://www.datacarpentry.org/spreadsheet-ecology-lesson/)*,* [*https://doi.org/10.5281/zenodo.570047*](https://doi.org/10.5281/zenodo.570047)

Data Carpentry develops and teaches workshops on the fundamental data skills needed to conduct research. Data Carpentry is a sibling organization of Software Carpentry. This material is extracted from Data Carpentry’s Lesson *Data Organization in Spreadsheets*.

*McNeil, J., Jones, M. A. (2018).*[*Data Management using NEON Small Mammal Data with Accompanying Lesson on Mark Recapture Analysis*](http://dx.doi.org/10.25334/Q4XH5S)*.*[*NEON - National Ecological Observatory Network*](https://qubeshub.org/groups/neon)*, QUBES.*[*doi:10.25334/Q4XH5S*](http://dx.doi.org/10.25334/Q4XH5S)

The National Ecological Observatory Network (NEON) is a program sponsored by the National Science Foundation and operated under cooperative agreement by Battelle Memorial Institute. This material is based in part upon work supported by the National Science Foundation through the NEON Program.

**The Data Sets**

The National Ecological Observatory Network:

* **NEON\_TreeVegetation\_variables.xlsx** – Metadata file for NEON terrestrial vegetation data describing the variable names.
* **NEON.Adaptation\_FacultyNotes\_DataManagementWithNEONTreeData.docx** – This document provides instructor information on the workshop activity.
* **NEON.Adaptation\_DataManagementWithNEONTreeData\_Slides.pptx** – This presentation accompanies the workshop activity and provides visuals to support the steps of the activity.
* **NEON\_tree\_data\_student.xlsx** –This workbook contains NEON terrestrial vegetation data consisting of both species identity and size (e.g., diameter at breast height) for all individuals at all NEON terrestrial sites.
* **NEON\_plot\_area.xlsx** - This workbook contains the area sampled at each of the sites which is needed to standardize the aboveground biomass data to cm per hectare.
* **NEON\_site\_climate.xlsx** - This workbook contains latitude, mean annual temperature and mean annual precipitation at each site.
* **NEON\_tree\_data\_instructor.xlsx** –This workbook contains NEON terrestrial vegetation data consisting of both species identity and size (e.g., diameter at breast height) already filtered by individuals and sites (according to the activity instructions) with pivot table result summaries and scatter plot figures.

**Instructor Notes on Student Instructions**

In this section, I lay out the instructor notes for the presentation. The information is designed to provide more context on specific sections of the activity. Also provided are some key points that students should consider for each discussion question. The slides were designed to be presented concurrently with the activity creating a discussion between instructor and students. Thus, students need to have access to as spreadsheet program like Microsoft Excel or Google Sheets in order to participate.

**Instructor’s Guide to the Lecture Presentation**

***What is Data Management and Why Is Important***

Presentation slide 3-6

This section opens the activity with a quick discussion on the **importance of data collection and management**. This needs to set up the question of **why scientists need to think about data management**. Data management is the practice of organizing and maintaining data in a convenient format for analysis.

***Spreadsheets and Dataset Format Explorations***

Presentation slides 7-9

This section is design to discuss with students **what is a spreadsheet** and expose them to the format of datasets from different research groups so that they can identify and understand commonalities in data organization and structure. A spreadsheet is an electronic document in which data is arranged in rows and columns of a grid. This format allows you to know exactly what variables were measured in the field and what is the sample size. In this way, **the format allows you to count, sort, and filter observations (rows) by variables (columns)** quickly and easily. In the Quick Exercise of slide 9, students can see that having two variables in one category (column) limits the number of independent variables in a statistical analysis.

If students have analyzed other datasets prior to this activity, you may use it as another example of dataset format. This could be more effective in the discussion as they would know its format and could relate it to how easy or challenging was to work with such data. After this discussion, student should be able to come up with the set of rules in slide 8.

***Formatting Data in Spreadsheets While Avoiding Common Errors***

Presentation slides 10-16

This section introduces techniques for **formatting data tables**. Instructors may separate students in groups of two for discussion of each exercise. In short time, students should be able to at least identify what is wrong with these spreadsheets, discuss the steps they would need to take to clean up using multiple tables and tabs, and generate a header for this new spreadsheet. The data management should yield a header with all available variables arranged in columns. This exercise introduces the following common spreadsheet errors demonstrated in slides 10-14. The terrestrial vegetation data file provided to the students do not include any of these errors.

***Quality Control***

Presentation slides 15-16

This section introduces techniques for quality assurance and control through **validation of data**, **data sorting**, and **conditional formatting**. With this exercise students will learn how to control for errors in data entry and how to identify them if they indeed occurred. This data management exercise should also yield a clear understanding of why common errors discussed (i.e., not filling in zeros, placing comments or units) should be avoided in spreadsheets.

***Putting Everything into Practice:***

***Managing a Large, Open Access Dataset to Answer a Research Question***

*Big data revolution and sharing*

Presentation slides 17-18

This section should be used to briefly discuss with students the **advantages of knowing good data management in a world of both big data revolution and data sharing**. The accumulation of data in the last decades, together with ethical revisions on data sharing, have made large amounts of scientific data accessible. These data could be used to test hypothesis and generate new questions. From this section to the end of the activity, instructors should highlight the **authenticity of research carried out with publicly available data**. Instructor should break with any misconception that students may have about the legitimacy of research when they do not generate the data themselves.

*NEON as a resource of open data*

Presentation slides 19-23

With the previous section as introduction, this part of the presentation should be used to introduce **NEON as an example of publicly available (open) data** to understand ecological processes. For a clear understanding of the ethics behind data sharing, the instructor should highlight who sponsors NEON (the National Science Foundation, a public agency built and administered with public funding). Instructors should also highlight vocabulary such as “open data” (free access) and discuss its meaning. Slide 21-23 gives an overview of NEON’s terrestrial vegetation structure protocol.

*Understanding shared data through metadata*

Presentation slides 24-37

When data is shared from an open source, data is often shared in a spreadsheet. Thus, researchers need clear knowledge on how the data was collected, as well as how the data was entered to the spreadsheet. This section can be used to discuss briefly **NEON’s data steps from the field to the spreadsheet**. A brief look at **NEON\_TreeVegetation\_variables** will help students realize the importance of documentation (i.e. providing metadata with spreadsheets) so that data can be used by researchers who did not collect the data. This metadata file describes the content, quality, condition, and other characteristics of the dataset.

*Data management*

Presentation slides 28-31

This section presets **guidance for managing NEON’s data** appropriately in order to answer the research question in an inquisitive-method style. Students should be able to manage the dataset and answer all questions.

**Question 1**: What is the *N* for each site?

* + 1. ***Hint***: Filter your data and create a new file
			1. siteID keep: BART, GRSM, HARV, LENO, MLBS, ORNL, SCBI, TREE, UNDE
			2. stemDiam remove < 3 values and NA
			3. scientificName remove: unknown
			4. growthForm keep: multi bole tree, single bole tree and small tree
			5. plantStatus keep: Live…., many categories of live (keep them all)

**Question 2**: What is the total biomass for each site?

* + 1. Calculate aboveground biomass
		2. Take the species-specific constants provided (column S and T; Chojnacky et al 2015) to calculate aboveground standing biomass in a new column.
		3. Allometric equation from (Chojnacky et al 2014\*)
			- **Ln (Biomass) = b0 + (b1\*ln(dbh)).**

**Question 3**: What is the total biomass per hectare for each site?

 3. Learn Pivot Tables:

* + 1. Click the appropriate field name and put into the values area.
		2. Change the value for your question.
			- What information from ID when it is counted would provide?
			- What value of biomass would be most appropriate to answer your question?

c) Don’t forget one of your questions is about biomass per hectare!

* + - * Need to open up the data file **NEON\_plot\_area** and obtain the area sampled per site in order to convert biomass at each site to biomass per hectare at each site.

**Question 4**: What is the relationship between total aboveground tree biomass and climate (choose either latitude, mean annual temperature, or mean annual precipitation)?

4. Create Figure:

* 1. Link data between biomass per hectare per site with a climate metric (e.g., latitude, mean annual temperature, or mean annual precipitation) at each site. Different climate metrics can be found in **NEON\_site\_climate** file.
	2. Choose the most appropriate type of graph for the data you have.
		+ - Determine which variable should be on the x and y axis
			- Remember to label axes and provide units if applicable
			- Title should always go on the bottom of the figure NOT on the top