**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 combines the *Data Carpentry: Data Organization in Spreadsheets Ecology Lesson* (Bahlai and Teal 2017) and the teaching module *Data Management Using NEON Small Mammal Data* (McNeil and Jones 2018) to address the second challenge by incorporating already developed spreadsheet data management exercises to answer an extended question on NEON small mammal data.

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**

*Data Carpentry: Data Organization in Spreadsheets Ecology Lesson* activity from Software Carpentry was added to McNeil and Jones (2018) in order to emphasize on proper data management when using large datasets. Specifically, this adaptation consists of three exercises that introduce students to 1) format spreadsheet data tables, 2) carry out spreadsheet quality control, and 3) count/sort/filter data of interest in order to conduct a pilot analysis on NEON small mammal abundance.

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.

**Original modules**

This lesson is 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**

Data Carpentry:

* **small\_mammal\_community.xls** – Subset of the small mammal data from southern Arizona addressing the effects of rodents and ants on the plant community. This file is a modified version of the original file from the Data Carpentry named survey\_data\_spreasheet\_messy-2.xls. The name was changed to better match this activity. This .xls contains two years of small mammal community data with multiple data table formats. This file should be used with students to identify common errors in formatting data tables and re-organized the data as recommended.
* **formatted\_small\_mammal\_comm.xls** – Subset of the small mammal data from southern Arizona. This file is the same as the original file from the Data Carpentry named survey\_sorting\_exercise.xls. The name was changed to better match this activity. This .xls contains combined years of small mammal trapping data with a correct structure of organization. This file should be used with students to learn how to carry out spreadsheets quality assurance and control.

The National Ecological Observatory Network:

* **NEON.D02.SCBI.DP1.10072.001\_variables.csv** – Metadata file for NEON small mammal data (DP1.10072.001) describing the variable names.
* **NEON.D02.SCBI.DP1.10072.001.readme.txt** – Metadata file for the NEON small mammal data (DP1.1072.001) providing more information on the data product.
* **NEON.FacultyNotesAdaptation\_DataManagementWithNEONSmallMammalData.docx** – This document provides instructor information on the workshop activity.
* **NEON.Adaptation\_DataManagementWithNEONSmallMammalData\_Slides.ppt** – This presentation accompanies the workshop activity and provides visuals to support the steps of the activity.
* **NEON\_small\_mammal\_data.xlsx** – This file is a modified version of NEON’s original file named NEONTeachingModule\_NEONSmallMammalDataAbundanceWorkbook.xlsx. This workbook contains only the September and October data from the NEON.D02.SCBI.DP1.10072.001.mam\_pertrapnight.072014to052015.csv file.

**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 3-5

This section opens the activity with a quick discussion on the **importance of data management** through its definition and the identification of format errors in a hypothetical data table. 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. A correct data table format is the first step of good data management using spreadsheets. In the Quick Exercise of slide 4, students can see that having two variables in one category (column) limits the number of independent variables in a statistical analysis.

***Spreadsheets and Dataset Format Explorations***

Presentation slides 6-9

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.

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 9.

***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 the 2013 and 2014 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 12-16. If time permits, Exercise 1 can be used to format appropriately the entire Arizona small mammal data set provided (as shown in slide 11).

***Quality Control***

Presentation slides 17-19

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 20-21

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 22-31

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 31 introduces NEON’s mark-recapture protocol on small mammals introducing Exercise 3 on answering a research question using NEON’s data. The number of NEON protocol-related slides in this section can be shortened as needed.

*Understanding shared data through metadata*

Presentation slides 32-34

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**. If students are expected to collect their own data at some point of the semester, spend the needed time discussing slide 32. This slide shows **NEON’s data collection sheet**. Students should be able to notice that the data collection sheet is organized into columns and rows, resembling a spreadsheet. Discuss with students why such an arrangement is helpful to complete the data process from the field to the spreadsheet.

Slide 33-34 introduces students to the research question in Exercise 3. This section should be used to demonstrate how to download NEON’s data and discuss the usefulness of **metadata files** when using open data to answer research questions. A metadata file describes the content, quality, condition, and other characteristics of a dataset. Instructors should note that the NEON website appearance will likely change over time and they may want to update screenshots in the presentation to match the current website.

*Data management*

Presentation slides 35-36

This section presets **guidance for managing NEON’s data** appropriately in order to answer the research question in an inquisitive-method style. It is highly recommended to discuss mark-recapture methods and Lincoln-Petersen estimator previous to this lecture in order to maintain this activity focused on data management. Students should be able to manage the dataset and answer all questions concluding that there are four elevation categories in the data, but only two with more than one capture occasion (September and October). Thus, their pilot analysis is limited to two elevation categories.

**Question 1**: What is the *N* for each elevation? ***Hint***: Use “Filter” and “Count”.

* + Filter by scientificName “*Peromyscus leucopus*”
  + Filter by elevation
  + Count each row per elevation

**Question 2**: Is there recapture data? What is the *N* of recaptures across elevations?

* + Filter by collectDate “*September*” and “October” (focus only in months)
  + Count each row per elevation per month

**Question 3**: Are elevation categories represented by different plots?

* + Filter by plotID

*Data analysis*

Presentation slides 37-39

This section is designed to **answer the research question** by estimating abundance using Lincoln-Petersen. Slide 37 encourages students to design a new analysis sheet with the estimator equation. With this exercise, students should realize that they can prepared a spreadsheet for analysis previous to adding any data to it. Slide 38 shows a possible way of preparing the spreadsheet with the estimator equation ready to be employed once students substitute the value of each variable in the equation. Highlight here that they are not managing the data anymore, they are analyzing it so creating a new sheet or adding the four variables of the Lincoln-Petersen estimator in the first column could be helpful for quick analysis. Students’ last task is to present their results in a figure. This exercise can be used to discuss the fact that, even that the independent variable is “continuous” to some extent, it can be treated as discrete (i.e., “low elevation”, “high elevation”).