**Module Overview**

Undergraduate STEM students are graduating into professions that require them to manage and work with data at many points of a data management life cycle. Within ecology, students are presented not only with many opportunities to collect data themselves, but increasingly to access and use public data collected by others. This activity introduces the basic concept of data management from the field through to data analysis. The accompanying presentation materials mention the importance of considering long-term data storage and data analysis using public data.

This data set is a subset of small mammal trapping data from the National Ecological Observatory Network (NEON). The accompanying lesson introduces students to proper data management practices including how data moves from collection to analysis. Students perform basic spreadsheet tasks to complete a Lincoln-Peterson mark-recapture calculation to estimate population size for a species of small mammal. Pairs of students will work on different sections of the datasets allowing for comparison between seasons or, if instructors download additional data, between sites and years. Data from six months at NEON’s Smithsonian Conservation Biology Institute (SCBI) field site are included in the materials download. Data from other years or locations can be downloaded directly from the NEON data portal to tailor the activity to a specific location or ecological topic.

In this activity, students will:

* discuss data management practices with the faculty. Presentation slides are provided to guide this discussion.
* view field collection data sheets to understand how organized data sheets can be constructed.
* design a spreadsheet data table for transcription of field collected data using good data management practices.
* view NEON small mammal trapping data to a) see a standardized spreadsheet data table and b) see what data are collected during NEON small mammal trapping.
* use Microsoft Excel or Google Sheets to conduct a simple Lincoln-Peterson Mark-Recapture analysis to estimate plot level species population abundance.

Please note that this lesson was developed while the NEON project was still in construction. There may be future changes to the format of collected and downloaded data. If using data directly from the NEON Data Portal instead of using the data sets accompanying this lesson, we recommend testing out the data each year prior to implementing this lesson in the classroom.

This module was originally taught starting with a field component where students accompanied NEON technicians during the small mammal trapping. As this is not a possibility for most courses, the initial part of the lesson has been modified to include optional videos that instructors can use to show how small mammal trapping is conducted. Instructors are also encouraged to bring small mammal traps and small mammal specimens into the classroom where available.

**The Data Sets**

 The National Ecological Observatory Network 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 following datasets are posted for educational purposes only. Data for research purposes should be obtained directly from the National Ecological Observatory Network (www.neonscience.org).

Data Citation: National Ecological Observatory Network. 2017. Data Product: NEON.DP1.10072.001. Provisional data downloaded from [http://data.neonscience.org](http://data.neonscience.org/). Battelle, Boulder, CO, USA

**Activity Data Sets**

* **NEON.D02.SCBI.DP1.10072.001.mam\_pertrapnight.072014to052015.csv** – Small mammal trapping data from NEON’s Smithsonian Conservation Biology Institute (SCBI) field site for July 2014 to May 2015 (seven months total as no trapping was conducted November to March). This .csv contains combined months from the combined small mammal trapping data from September and October 2014. This data are a combined table based on the individual site-month tables available from the NEON Data Portal. This file should be used with a group of student to implement their skills and allow different partners to estimate abundance for different plots and seasons to allow for larger patterns to be examined in the classroom.
* **NEON.D02.SCBI.DP1.10072.001\_variables.csv** – Metadata file for NEON small mammal data (DP1.10072.001) describing the variable names.
* **NEON.DP1.10072.001\_readme.txt** – Metadata file for the NEON small mammal data (DP1.1072.001) providing more information on the data product.
* **NEON.D02.SCBI.DP1.10072.001.20140917-20140924.xml** – Example XML file for the September 2014 data that is provided with this lesson.
* **Abbreviated NEON Small Mammal Trapping Protocols.pdf** – This document is adapted from NEON Document NEON.DOC.000481 in May 2017. It describes the layout of NEON small mammal sampling grids and is only intended for use with the educational teaching module outlined in this article.
* **NEONSmallMammal\_SCBI\_BlankDataSheet.pdf** – This document shows how a NEON field data sheet is formatted and is provided to show a potential format for recording this type of data. This document is limited to D02 data sheets, all NEON Small Mammal trapping data sheet can be found in document *NEON.DOC.001585 Datasheets for TOS Protocol and Procedure: Small Mammal Sampling*.

**Related Documents**

* **NEONTeachingModule\_Teacher\_DataManagementWithNEONSmallMammalDataAndMarkRecapture .docx** – This document. Provides additional instructor information on the activity.
* **NEONTeachingModule\_Student\_DataManagementWithNEONSmallMammalDataAndMarkRecapture .docx** – Can be used as a handout or guide for the students to progress through the activity. The instructor will need to make a few modifications to fit the specific teaching setting.
* **NEONTeachingModule\_DataManagement\_Slides.ppt** – This presentation accompanies the student activity and provides visuals to support the steps of the activity.
* **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. Each worksheet lays out one step needed to conduct that Lincoln-Peterson species population abundance analysis.

**Instructor Notes on Student Instructions**

In this section we layout the student instructions (provided in a separate document for ease of editing) accompanied by instructor notes. The information is designed to provide more context on specific sections of the the lesson. Also provided are some key points that students should consider for each discussion question.

The presentation slides included with the activity (*NEONTeachingModule\_DataManagement\_Slides.pptx*) can be used to highlight key points as the students work through each step of the activity. Below we’ve indicated which slides correspond to which portion of the activity below. They are meant as starting point and framework for faculty to tailor the activity to their own classroom. The slides are designed to be presented before the students start on the activity, but could also be used concurrently with it.

***Before class:***

Review the following resources to prepare for the data management activity:

* NEON website:<http://www.neonscience.org/>
* Abbreviated NEON Small Mammal Trapping Protocols
* Thibault, K. NEON breeding bird and small mammal abundance and diversity sampling. NEON: <http://data.neonscience.org/api/v0/documents/FSUHandout_Vertebrates.pdf>
* *Instructor Note: Select a video to show in class or have students watch prior to class starting*
	+ National Park Service. From Field to Lab: Small Mammal Monitoring in Denali National Park: <https://youtu.be/KvGvS8pApFE> (1:32 - 2:30 highlights small mammal trapping/handling techniques)
	+ University of Oxford. The Laboratory with Leaves (Part 10): Small Mammals: <https://youtu.be/bIjva3pa2YA> (This video provides context for why small mammal monitoring is important to ecology in general).
* Sutter, R.D., Wainscott, S.B., Boetsch, J.R., Palmer, C.J. and Rugg, D.J. (2015). Practical guidance for integrating data management into long-term ecological monitoring projects. Wildlife Society Bulletin, 39: 451–463.
* Borer, E.T., Seabloom, E.W., Jones, M.B., and Schildhauer, M. (2009). Some simple guidelines for data management. Bulletin of the Ecological Society of America, 90(2): 205–214.

*Make sure to bring your laptop to class/meet in the computer lab on [insert date of class session]* and make sure *Microsoft Excel/Google Sheets* is installed. You’ll be working in partners to enter and analyze data that day.

***Why Data Management is Important***

*Instructor*

Presentation slides 2-7

This section is primarily covered through the pre-course student readings and through the slides. This needs to set up the imperative of why scientists need to think about data management. If your students are not motivated by field ecology examples, there are many situations and other lines of work where data management is still essential. If you feel this might be the case with your students, have them brainstorm situations where they’ve encountered or can imagine good and bad data management strategies in their preferred line of work.

*Student*

Data are the backbone of scientific research and exploration, so learning how to collect and process data efficiently is a critical skill for scientific professionals. Most people are not born understanding how collect, record, enter, and analyze data, but with guidance and practice you can learn how to handle information and create world-class datasets.

Scientific organizations, especially large ones, spend a lot of time and effort determining the best ways to process data. The National Ecological Observatory Network (NEON) is one such organization that has made efficient data collection and processing a priority. NEON was designed to collect long-term ecological data on a continental-scale to help researchers address questions related to climate change, land-use, and invasive species. Data are collected at field sites called domains using standardized protocols, which allow for comparisons across large geographic ranges. Data on dozens of different variables and species will be collected every year for 30 years, yielding a comprehensive look at ecological processes across the entire United States. Regardless of the variables being measured, the general flow for data in these projects progresses from data collection to data files and metadata files as shown in Figure 1.



Figure 1. A workflow from data collection (can be in the field, lab, or other venue) to data collection sheets (paper, app, or entry form) to data files and metadata files.

Given the scale and scope of the project, they will create literally terabytes of data every year, so the information needs to be very well organized to be useful. In this activity, you’ll get practice translating field data into a usable format for long-term archiving and then explore how real NEON data can be used to detect ecological patterns, in this case the change in small mammal abundance over a year.

***Field Collection & Data Sheets***

*Instructor*

Presentation slides 9-10

If resources permit, this section could be enhanced by having students actively participate in small mammal sampling. At the Smithsonian-Mason School of Conservation, we are fortunate to be located adjacent to a core NEON field site (SCBI), so students are able to observe NEON field technicians gathering this data in the field. Alternatively, if there are natural history specimens available for the target species in your area, bringing those into the class as visuals would also be a great addition to this portion of the activity. Finally, to save paper, it could be possible to have the students examine an electronic copy of the blank data sheet.

*Student*

Small mammals were chosen by NEON to be bioindicators because they are present across the country in a wide variety of habitats. Their small size and short lifespan makes them sensitive to environmental changes, and they are responsible for spreading or maintaining a wide diversity of zoonotic diseases in an environment. They are also easy to safely collect as live specimens using arrays of traps like those described in the *Abbreviated NEON Small Mammal Trapping Protocol.* Live trapping has the advantage of being able to return the animal to their habitat without having to destructively sample. As you learned in the readings and the Youtube videos, in just a few minutes you can collect a lot of information from an individual animal. Because researchers want to reduce the stress on the animal while it is captured, it’s important to have an efficient framework for recording that data.

Take a few minutes now and review the NEON Small Mammal sampling datasheet. See if you can identify what variable is being recorded in each of the column categories. Make sure you know what codes refer to what type of animal being collected

Now look at the example data sheet that is filled out. On the data sheet, circle the column headings for the following variables:

* Plot ID
* Date of capture
* Species
* Individual ID
* Sex
* Weight
* Whether individual is a recapture?

***Data Sheets & Data Files***

*Instructor*

Presentation slides 9-18

As the students work on creating their data tables (spreadsheets), the faculty should move between pairs and assist, but students should be allowed to struggle a little with interpreting the data sheet and developing an efficient way of entering the data. Students most often ask for clarification on the meaning of codes on the data sheet and for advice on the details of best practices for data organization. Giving them the space to experiment will help lead to a better discussion. The discussion questions can be addressed in the classroom or used as part of a written evaluation of the activity.

*Student*

Processing raw data sheets into a data table is only easy if the data table is well designed. Thinking about the presentation and the principles described in Borer et al. 2010 and Sutter et al. 2015, work in pairs to create an Excel data file that displays the information from your example data sheet for the variables you identified above. Make sure that your data table adheres to the best practices for data file construction that we talked about.

*Key points for discussion questions:*

1. Do you think the NEON data sheets are well designed to transfer the information to a data file? What makes the process easier and what makes it challenging?
	1. This will primarily be a matter of opinion. The data is laid out in a logical manner but some students may think that some field are too small or other similar comments.
2. Imagine you were responsible for entering data from hundreds of data sheets. How would you make sure you were not making mistakes? What types of checks could you do to make sure you were correctly transferring the data?
	1. Be sure students consider both manual checks (two people enter data, proof data that is entered, standard ways of writing numbers, layout similar to data sheet to avoid jumping around) and automated checks (set bounds – min or max values – on entry possibilities in forms).

***Public Data & NEON***

*Instructor*

**Presentation slide 21- 24**

This version of the activity is built around a pre-downloaded data set from NEON’s SCBI site, but if time and experience permits, students (or instructors) could download separate data sets directly from the NEON data portal from sites that are closer to their location. Accessing this information does require a good internet connection and not all field sites have all of their data available yet. Faculty interested in doing this should download data sets before class to make sure they are amenable to the analysis portion of the activity.

*Student*

Another hallmark of NEON is that the data are all publically available. NEON has created an online data portal (<http://data.neonscience.org>) that allows access to all of the NEON data from any Domain across the country. This portal will serve as the long-term repository and clearinghouse for all of the NEON data in perpetuity.

We will use a series of data files downloaded from this portal to estimate the abundance of several small mammal species in different seasons (spring, summer, and fall) at NEON’s Smithsonian Conservation Biological Station field site during 2014 and 2015.

***Metadata***

*Instructor*

Presentation slides 19-20

NEON uses the Ecological Metadata Language (EML) to provide metadata with all data downloads. In the example dataset (or in datasets downloaded by students), there is an accompanying .xml file. This is the metadata file. Instructors could choose to go into these files more, either as a demo, or if time and materials are available, the students could explore it.

The discussion on metadata can be expanded greatly to include many other NEON documents including the xml files and other documentation that can be downloaded with the data (see NEONDataDownloaded directory). These additional documents reflect many aspects of the site characteristics, sampling effort, and data processing. While most undergraduates are not involved in data collection at a scale where metadata are essential, being familiar with the concept can put them a step above the competition when they are applying to jobs or internships. Several students in the Smithsonian-Mason programs have commented that data management, including work with metadata, was expected of them in government and NGO internships they worked with after graduating.

*Student*

As we talked about during the presentation, metadata are another important component of collecting good data. A good metadata file can help someone unfamiliar with a data file interpret the codes and variables presented – and will help you remember what you did when you come back to the data later. It also provides an opportunity to discuss any irregularities in the data set.

*Key points for Discussion Questions:*

1. Examine the metadata file for the NEON data file. Briefly discuss with your partner how this file could have helped you interpret the data sheet and create your own data file or perform data analysis. Be prepared to share your observations with the class.
	1. The XML metadata file provides information that might be important for interpreting the results such as geographic location or elevation. It also provides information about who produced the data, how the data can be used, and how the data should be cited when used. This file is in a machine readable format that can be used with a scripted analysis.
	2. Also provided are the *readme* and *variables* files. These metadata files also provide some or all of the above information but are more human readable (but not machine readable).
	3. The *Abbreviated NEON Small Mammal Trapping Protocols* could also be considered a metadata file as it provides more information on how the data were collected.

***Data Analysis***

*Instructor*

Presentation slides 25-26

Faculty should take the time to clearly explain the formula and assumptions of the Lincoln-Peterson model before having the students engage with the data file. Having a clear idea of the type of data they are looking for will make it easier for them to perform the data analysis. Also faculty should assign partner groups for each section of the activity. As with the creation of the data file, faculty should circulate among pairs and provide advice and answer questions, but it is best if the students are able to explore the issues surrounding the data analysis themselves. As above, the discussion questions can be addressed in the classroom or used as a written evaluation.

One important assumption is made in the calculation of the Lincoln-Peterson method using this data set. NEON small mammal trapping data includes both pathogen plots (three night bouts) and diversity plots (one night bouts). For simplicity in the lesson, all nights of trapping in pathogen plots are lumped into a single bout for quantifying the variables in the Lincoln-Peterson method of population analysis.

Because there is so much concurrent data collected at NEON sites, this activity is also an excellent jumping off point for more sophisticated or alternative analyses. We’ve outlined a basic example of how these data could be used for analysis, but there are many ways the activity could be extended. For example, students could investigate community composition at different sites, or even the same plot. Pairs of students could work on data from different field sites to do a multi-site comparison. NEON terrestrial observation plots are laid out in a stratified random layout based on National Land Cover Database (NLCD) classes present at a field site. By incorporating NLCD data with the small mammal data, students can ask questions like how does the species composition in different NLCD classes vary or how does the abundance of a given species vary? Pairs of students could calculate abundance and determine the NLCD class(es?) for each plot. As background for this extension, faculty should introduce the concept of NLCD classes.

*Student*

Once you have a well-designed data file, you can use that information to determine interesting patterns. One of the simplest ways that the NEON small mammal datasets can be used is to calculate abundance estimates for individual species within the plots. There are many ways to estimate abundance. One of the simplest is the Lincoln-Peterson method. This calculation uses data about the recapture of marked individuals of a species to estimate how many individuals of that species are present in a particular habitat. Because NEON small mammal protocols include marking individual animals with unique numerical identifiers, we can easily use NEON datasets to calculate small mammal abundance using Lincoln-Peterson according the following equation.

$$\frac{n\_{1}}{N}=\frac{n\_{2}}{m\_{2}}$$

N = total population size estimate

n1 = Number of individuals captured and marked in first sampling bout

n2 = Number of individuals captured in second sampling bout

m2 = Number of marked individuals in second sampling bout

It’s important to note that there are several assumptions that should be met for this calculation to generate an accurate estimate of population size:

* Individuals are randomly distributed between captures
* There is no change in the population (i.e. births, deaths, immigration, emigration) between sampling bouts
* Marking individuals does not impact their likelihood of being captured again in the future

*Key points for discussion questions:*

1. Lincoln-Peterson estimation depends on several assumptions about the population. Knowing what you do based on the sampling methods outlined in the *Abbreviated NEON Small Mammal Trapping Protocols* document, do you think any of those assumptions have been violated in this data set? Why and what could be done to address those issues?
	1. As mentioned above, we are using data from the first month as bout 1 (n1) and from the second month as bout 2 (n2 and m2). It is likely some animals die or leave the plot between these two data points violating the assumption that there is no change in the population between sampling bouts.
	2. Some individuals are known to be “trap happy” or “trap shy” after being captured, and therefore, they may violate the assumption that marking individuals does not impact their likelihood of being captured again in the future.

Working in pairs, use the workbook – NEONSmallMammalDataAbundanceWorkbook.xlsx – as a guide to calculate the Lincoln-Peterson estimation of population abundance using the following protocol:

1. Open the data file (NEON.D02.SCBI.DP1.10072.001.mam\_pertrapnight.072014to052015.csv) using Excel.
2. You and your partner will perform the analysis for sampling bouts either in the spring (April to May 2015), the summer (July to August 2014), or the fall (September to October 2014) for samples collected in your assigned plot. Record your time frame and plotID below:

Time frame:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Plot ID:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Sort the data by plot ID and then by collectDate. Now you can see when trapping occurred at each plot.
2. You will perform the Lincoln-Peterson calculation for white-footed mice (*Peromyscus leucopus*, PELE). Therefore, filter your data for the specified taxonID and plotID. Now you see only the data of interest for you Lincoln-Peterson calculation.
3. Identify unique individuals collected throughout your time frame. Record the following:

Number of individuals captured during the first month of the time frame (n1): \_\_\_\_\_\_\_\_\_

Number of individuals captured during the second month of the time frame (n2):\_\_\_\_\_\_\_

Number of individuals from the first month recaptured during the second month (m2):\_\_\_

1. Use these numbers to calculate the population abundance of PELE for your site and time frame:
2. Share your results with the class.

*Key points for discussion questions:*

1. Based on everyone’s data, how does the population abundance change for white-footed mice between plots? What are some hypotheses for why this pattern may exist?
	1. Responses may depend on what time frames and plots are used. Hypotheses should include the difference in habitat between plots. The nlcdClass variable does provide some broad categories of different habitat that the plots are within. NEON terrestrial plots have a stratified random sampling scheme to match the NLCD classes at a field site.
2. Based on everyone’s data, how does the population abundance change for white-footed mice over the year at this site? What are some hypotheses for why this pattern may exist?
	1. Responses may depend on what time frames and plots are used. Hypotheses should include how the reproductive cycle may influence population size.

**Potential Challenges**

Possible solutions to challenges that other have encountered while implementing this lesson are detailed below.

1. Students do not have laptops to work with the files on their own equipment.

Solution: Prepare a computer lab with the required files before the class session or use course management software to upload the activity files so they can be downloaded in a computer lab during the exercise.

1. Students are unfamiliar with how to use spreadsheets, especially the sort and filter functions. Note that the activity is written referring to Excel, however, it can easily be adapted to Google Sheets.

Solution:

Option 1: Provide a pre-lab on the use of spreadsheets (Excel or Google Sheets). The HHMI Biointeractive Spreadsheet Data Analysis Tutorials (<http://www.hhmi.org/biointeractive/spreadsheet-data-analysis-tutorials>) may be useful for this.

Option 2: Pair students based on their experience such that students with Excel experience are matched with those who do not have it. If there are not enough students with experience, faculty can walk the class through the sorting and filtering steps as a demonstration or prepare pre-sorted and filtered versions of the data set for students to use.

Option 3: Use the *NEON Small Mammal Data Abundance Workbook* to demonstrate sorting and filtering functions to the class prior to the pairs working with the larger dataset to calculate their species abundances.

1. Students are uncomfortable with the concept of trapping small mammals.

Solution: Use the activity as a way to discuss Animal Care and Use standards and emphasize that all protocols are designed to minimize the impact and stress on the animals. Reinforce that responsible scientific sampling goes through review and is approved before being implemented. Also emphasize that there are not alternative methodologies for collecting this information; if we are to understand the biology and ecology of small animals like small mammals, they must be captured and handled.