TIEE

Teaching Issues and Experiments in Ecology - Volume 19, May 2023

ISSUES: DATA SET

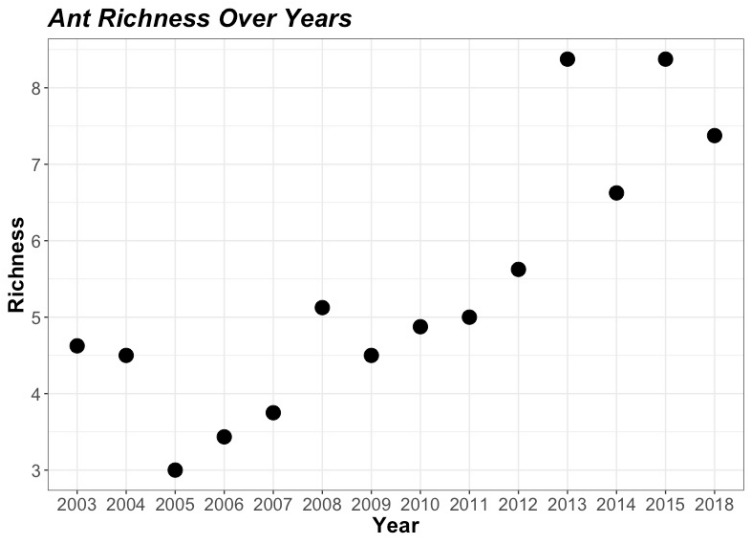
Playing with open biodiversity datasets: case studies using data from NEON, EDI, and GoMRI

Xuan Chen[1], Daijiang Li[2]

[1]Department of Biological Sciences, Salisbury University, Salisbury, MD 21801

[2]Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803

Corresponding author: Xuan Chen (XXCHEN@salisbury.edu)



Ant species richness over time at Harvard Forest between 2003 and 2018

THE ECOLOGICAL QUESTIONS:
How does species richness change across large spatial scales? How does biodiversity in one place change over time? How does human disturbance affect biodiversity?

FOUR DIMENSIONAL ECOLOGY EDUCATION (4DEE) FRAMEWORK

- **Core Ecological Concepts:**
  - Community
    - Species diversity – Biodiversity
- **Ecology Practices:**
  - Quantitative reasoning and computational thinking
    - Data skills – inputting and data-mining /data visualization
    - Computer skills: spreadsheets, "R"
- **Human-Environment Interactions:**
  - Human accelerated environmental change
    - Anthropogenic impacts
- **Cross-cutting Themes:**
  - Spatial & Temporal
    - Scales

WHAT STUDENTS DO:
Use R and an open biodiversity dataset to explore ecological questions related to the effects of space, time, and human disturbance on biodiversity

STUDENT-ACTIVE APPROACHES:
In this module, two to four students work collaboratively in a group to complete assignments. The module incorporates mostly open-ended questions, which can be challenging to answer due to the overwhelming amount of information available on the internet. Additionally, writing R code as an undergraduate can be a daunting task due to the complexity of the coding process and lack of experience. Collaborative learning can help to reduce the stress and anxiety associated with these challenges and create a more supportive learning environment. The assignments are uniform across all groups, with each group submitting one answer sheet to the instructor.

STUDENT ASSESSMENTS:
Answer sheet including questions' answers of each activity, R scripts (demonstrating how students manipulate the code for the formative and summative assessments), figures and tables related to analysis results.

CLASS TIME:
This module was designed to let students finish the exercises within two 3-hour labs: Lab 1: Activity 1 -> Introduce R -> Activity 2 Lab 2: Activity 3 -> Activity 4 -> Summative assessment

COURSE CONTEXT:
upper-level courses (e.g., General Ecology, Community Ecology, Conservation Biology) in Biology and related majors (e.g., Environmental Science). The best time to introduce this module is during the second half of a course when students have learned the concept of species, diversity, resilience etc.

SOURCES:

References

Ellison, A. (2021). Ant Assemblages in Hemlock Removal Experiment at Harvard Forest since 2003 (Reformatted to the ecocomDP Design Pattern) ver 5. Environmental Data Initiative. https://doi.org/10.6073/pasta/cd33e1651c73841976049a5730504ad1

Gaynor, S. (2020). Introduction to R with Biodiversity Data. Biodiversity Literacy in Undergraduate Education, QUBES Educational Resources. https://doi.org/10.25334/84FC-TE88

Li, D., S. Record, E.R. Sokol, M.E. Bitters, M.Y. Chen, A.Y. Chung, M.R. Helmus, et al. (2022). Standardized NEON Organismal Data for Biodiversity Research. Ecosphere 13(7): e4141. https://doi.org/10.1002/ecs2.4141

Rabalais, N. (2016). Above ground plant biomass, canopy height and estimated percent cover supporting marsh and subtidal benthic community and marsh invertebrate distribution studies in paired oiled/unoiled sites in coastal Louisiana in Spring and Fall 2013. Distributed by: Gulf of Mexico Research Initiative Information and Data Cooperative (GRIIDC), Harte Research Institute, Texas A&M University–Corpus Christi. https://doi.org/10.7266/N7X34VD2

Data sources

https://www.neonscience.org/data-collection/ground-beetles

https://portal.edirepository.org/nis/mapbrowse?packageid=edi.193.5

https://data.gulfresearchinitiative.org/data/R1.x139.144:0023

ACKNOWLEDGEMENTS:
This study was funded by National Science Foundation (**DBI-1730526** RCN-UBE: Biodiversity Literacy in Undergraduate Education - Data Initiative and **DBI-2120678** RCN-UBE: Transforming Ecology Education to Four Dimensional Network). We would like to thank Anna Monfils and Luanna Prevost for their help in developing the module. We also thank reviewer and editor for their insightful comments.

OVERVIEW OF THE ECOLOGICAL BACKGROUND

Working in the field, getting your hands dirty, and appreciating the natural scenery are all fun parts of doing outdoor ecological research. While analyzing data collected by yourself (primary data) can be used to answer ecological questions at local spatial or short temporal scales, ecologists nowadays also often use existing large data to generate research questions and test hypotheses, especially when the studies are about large spatial scales and long temporal scales (e.g., changes of global biodiversity in the last several decades). Many organizations provide free, online access to large ecological datasets.

In this module, we will "play" with open ecological datasets from the internet. First, we will learn about some websites that host open diversity data. Then, we are going to investigate: (1) latitudinal gradients of biodiversity using National Ecological Observatory Network (NEON) data; (2) long-term changes in biodiversity using data from the Environmental Data Initiative (EDI); and (3) the effects of anthropogenic disturbances on biodiversity using data from Gulf of Mexico Research Initiative (GoMRI). The instructor will show you how to use R for data manipulation, analysis, and virtualization. (Don't worry about whether you are familiar with R at this moment, this module is not primarily about programming. The instructor will introduce the basic functions of R. You don't need to have experience writing R code.)

LEARNING OBJECTIVES:

Upon completion of this module, each student should be able to:

- Explain the concept and importance of biodiversity and open data sources
- Download, clean, and organize open biodiversity data
- Perform basic statistical analysis of biodiversity data and visualize results
- Use R and relevant packages for data organization and analysis
- Be aware of that ecological processes and patterns may shift across temporal and spatial scales
- Become familiar with the impacts of oil pollution on ecosystems
- Identify appropriate open ecological datasets when addressing a question (optional)

DATA SETS

- NEON.R: R script for download and analysis of NEON data.
- EDI.R: R script for download and analysis of EDI data.

- GoMRI.R: R script for analysis of GoMRI data.
- GoMRI_RabalaisPlantData.xlsx: Data for GoMRI analysis.

STUDENT INSTRUCTIONS

First, let's form study groups consisting of 2-4 students per group. You can begin by independently seeking answers to the questions before coming together to discuss and exchange insights with your group mates.

ACTIVITY 1: OPEN BIODIVERSITY DATA

The aim of this activity is to enhance your self-learning and critical thinking skills by challenging you to utilize Google and other search engines to find answers to specific questions. The activity is designed to teach you how to navigate vast amounts of information available on the internet, enabling you to locate the knowledge you need with greater ease and efficiency. Additionally, you will learn how to evaluate the credibility and reliability of sources, thereby developing critical thinking skills essential for ecological research.

Question:

1. What is open data? Why is open data important for scientific research?

2. What is biodiversity (hint: biodiversity is not just the number of species)? What is biodiversity data?

Formative assessment: find one peer-reviewed ecological paper that used open data and summarize that paper in no more than five sentences. Include the reference to the paper.

TIEE

Teaching Issues and Experiments in Ecology - Volume 19, May 2023

ACTIVITY 2: NEON (https://www.neonscience.org)

The aim of this activity is to familiarize you with the National Ecological Observatory Network (NEON) and use example data from NEON to examine changes in beetle diversity across broad spatial scales.

Question:

1. What is the National Ecological Observatory Network (NEON)? What are its goals? Why is NEON important for ecological research?

2. How many sites does NEON have? What ecoclimatic domains do those sites cover?

3. What type of environmental (abiotic) data does the NEON project generate in terrestrial and aquatic sites?

4. What type of biodiversity data does the NEON project generate in terrestrial and aquatic sites? How many years of data are available?

5. If you have the opportunity to add an organism to the NEON project, which organism would you choose and why?

6. What is meant by Latitudinal Gradients of Biodiversity?

For the next two questions, we are going to consider NEON data on beetle diversity in the US as an example of how open data can be used to address ecological questions at broad spatial scales. The R script will import the data from the NEON database and lead you through addressing the questions.

7. Follow the instructor's direction to address the following questions using R: (1) how many degrees of latitude and longitude do those sampling sites cover? (2) Describe the pattern of the beetle diversity across the latitude in the U.S. (3) Is this pattern consistent with the latitudinal gradients of biodiversity? If the pattern is different from what is expected, what might explain the difference?

8. Separate the dataset into three parts (east, middle, and west) based on the longitudes of the study sites. Repeat the analysis for each part. Is the pattern similar among the three regions? If not, how does the pattern differ? What might explain any regional differences in biodiversity gradients?

Formative assessment: modify the R script and study the latitudinal gradient of bird using the NEON data. Do the diversity of bird and beetle show similar latitude gradients? If not, can you think about some explanations?

TIEE

Teaching Issues and Experiments in Ecology - Volume 19, May 2023

ACTIVITY 3: ENVIRONMENTAL DATA INITIATIVE (HTTPS://EDIREPOSITORY.ORG)

The aim of this activity is for you to learn about the Environmental Data Initiative (EDI) and then use data from EDI on ant diversity at Harvard Forest to explore changes in ant diversity over time.

Question:

1. What is the mission of the Environmental Data Initiative (EDI)?

2. What types of data does EDI hold? What are the differences between NEON and EDI?

For the next two questions, we will investigate ant diversity over time at Harvard Forest as an example of how open data can be used to explore ecological questions across broad temporal scales. The R script will import the data from the EDI database and lead you through addressing the questions.

3. Follow the instructor's direction to address the following questions using R: How did ant diversity at Harvard Forest changed over time?

4. Repeat the analysis of the above question, but this time only use data from 2003-2009, as an example of a shorter timeframe for analysis. Are you seeing the same pattern? If not, how does it differ, and why might the pattern have changed?

Formative assessment: modify the R script and study the change of heron between 1992 and 2011 on Chincoteague Island based on this dataset (id = "edi.324.3") Are the temporal trends that you found for ants at Harvard Forest similar to those that you found for herons on Chincoteague Island? What might explain any similarities or differences?

TIEE

Teaching Issues and Experiments in Ecology - Volume 19, May 2023

ACTIVITY 4: DEEPWATER HORIZON OIL SPILL

The last dataset that we will explore is from studies of the ecological impact of the Deepwater Horizon oil spill.

Question:

1. Find some examples about how the Deepwater Horizon oil spill affects coastal ecosystems (Google it).

2. Study this website (https://gulfresearchinitiative.org) and answer (1) what is GoMRI? (2) What type of data does the website hold?

For the next two questions, we will consider plant diversity data from marshes impacted by the Deepwater Horizon Oil Spill. Before answering the questions, download the data from https://data.gulfresearchinitiative.org/data/R1.x139.144:0023.

1. The researchers collected the plant biomass data at the edge of the marsh (1 meter from the edge) and inside the marsh (10 meters from the edge). Why might the research have designed the sampling procedure in this manner?

2. Using the R script provided, determine if oil pollution reduced plant biomass at the marsh edge (1 meter from the edge).

*Formative assessment: modify the R script and study if the oil affected the plant biomass **inside** the marsh (10 meters from the edge). Is the effect of the oil spill on plant biomass the same inside and at the edge of the marsh? What might explain any differences you might see.*

SUMMATIVE ASSESSMENT (OPTIONAL)

Think about an ecological topic that interests you (e.g., invasive species, climate change), develop a hypothesis (e.g., invasive species would decrease local diversity), and search for an open dataset that can test your hypothesis. Alternatively, return to the R code that you used in the previous activities. The code shows you what variables are available in each dataset. Based on the data available in a particular dataset, develop a hypothesis and then use the data available in the dataset to test your hypothesis.

Some other useful open data resources and R packages

https://www.idigbio.org
http://idigbio.github.io/2015-10-05-PrePRAGMA/02-ridigbio.html

https://www.gbif.org
https://www.r-bloggers.com/2021/03/downloading-and-cleaning-gbif-data-with-r/
https://poldham.github.io/abs/gbif.html

TIEE

Teaching Issues and Experiments in Ecology - Volume 19, May 2023

NOTES TO FACULTY

THIS IS NOT A R PROGRAMMING MODULE! It is not necessary for students to know how to use R before the class. The purpose of this module is to show students the power of open data. However, knowing the basics of R will facilitate the learning process. Thus, it will be great if the students can finish the module of Gaynor 2020 (doi:10.25334/84FC-TE88) before the class starts. Or the instructor can spend some time to explain the R script of Activity 2. Again, students don't need to understand every detail of the script.

Making groups will facilitate the learning process because: (1) some questions ask students to find the answers by themselves through using search engines, discussion within the group can speed up the process. (2) Students can help each other during coding (e.g., checking syntax errors). It would be great if each group contains one student who has some experience of programming.

Prerequisite (student):

Before starting this module, students should

- Know how to search papers in the digital database (e.g., Google Scholar, Web of Science)
- Have R and RStudio installed in the computer

Prerequisite (instructor):

Instructors should be familiar with the R environment and basic coding.

Comments on Student Assessments:

Answers of the first two questions of each activity are easy to find on the internet. Students just need to click the website links or type the keywords in search engines. For example:
Activity 1.1: https://opendatahandbook.org/guide/en/what-is-open-data/
https://sco.library.emory.edu/research-data-management/open-data/benefits-research.html

Activity 4.1: https://www.natlenvtrainers.com/blog/article/the-environmental-impact-of-the-deepwater-horizon-oil-spill

In Activity 2, instructors can choose the taxonomic group with which they are most familiar with. Instructors can also encourage students to explore the data from multiple years and see if the patterns vary among years. In addition, linear model is introduced here to study the correlation between diversity and latitude. However, it's worth noting that while a linear model can provide some insights, it may not accurately capture the complexity of the relationship for all taxonomic groups. Instructors who wish to delve deeper into this topic may consider discussing alternative models such as polynomial or additive models.

Answer keys for formative assessment of each activity:

- Activity 2: change line 14 to bird_raw <- neonDivData::data_bird, the data name should also be modified in the rest of the script.
- Activity 3: change line 16 to data_full <- read_data(id ="edi.324.3"), the data name should also be modified in the rest of the script.
- Activity 4: change line 47 to data_full_10 <- subset(data_full, distance_edge == 10) , the data name should also be modified in the rest of the script.

Some questions are open-ended. For instance, the species distribution patterns may not be consistent with ecological hypotheses and theories. This provides a good opportunity to ask students some further questions like: (1) is the dataset suitable for testing the hypothesis? (2) Do we expect species from different taxonomic groups to show similar distribution patterns in nature? (3) Will different functional groups (generalist vs. specialist species) respond the same to disturbances? (4) Should we separate invasive and native species when we perform the analysis? The type of those further questions depends on the level of the course. These questions could be used as prompts during discussions of the answers. Alternatively, the questions could be included in a revised student handout that provides additional prompting and structure for students in introductory courses.

Answer keys for summative assessment: Open data used in this module activities are easy to access, and the data manipulation is not that hard. However, many open datasets are messy and not easy to explore, especially using R. The best practice is probably to let the students explore the data in NEON or EDI and do a little data mining. Note that this summative assessment can be very challenging for students who don't have coding experience. Thus, this assessment is optional