LTER Ecology Data Archives Educate Tomorrow's Scientists

Testing Metadata, Reproducing Results, & Training Students for Synthesis Science

Judith Cushing and Kathleen Saul
The Evergreen State College, Olympia, WA

Motivation

1. Are published results that accompany ecology archives replicable? If not, would an inability to verify published results of scientific research "lead to a credibility crisis affecting … scientific fields" (Symposium on Reproducibility and Interdisciplinary Knowledge Transfer, V. C. Stodden et al, AAAS, February, 2011)?
2. Can using LTER data archives effectively educate students to use ecology data archives, analyze real world data, and conduct synthesis science?
3. Which LTER metadata are useful to data users, and how might metadata be improved?

Project Assignment

In Evergreen's 2011 Master's of Environmental Studies Quantitative Methods course, 26 students worked in 12 teams to analyze one or more data sets from the H.J.A. LTER, using both metadata and published results. Students conducted new analyses or tried to reproduce reported results using R. The project included a field trip to HJA . Students were asked to articulate their process for managing data, what they learned about statistical analysis using existing data sets, what they found most difficult, and how to improve metadata.

HJA datasets used were: AS006; DF014 (1,3,5); GSWS01,2,3; HF07; SA015; SA021; SP002 with EVMP; TD014, TD017, TD018, TD021; TD035;TP114; TSBR, Ecotone 7, 8, 9, 10; and unpublished data from seven different temperature sensors.

Analysis Process*

1. Identify topic; find candidate data set(s)
2. Conduct background research
3. Visit field site; refine topic and choose data set(s)
4. Articulate research question(s)
5. Download data set(s) of interest, study metadata
6. Run exploratory analyses (descriptive statistics)
7. Develop statistical hypotheses, choose statistical tests
8. Clean or transform data and rerun exploratory analyses
9. Perform analyses
10. Write report and present work

Advice from Students:

******* Don't jump into analysis before exploring and cleaning the data. Spend time with the raw data, do lots of descriptive analyses, try to grasp the whole picture first. Start in Excel and then move to R! Be patient: real data take longer than a "cooked" dataset. Analysis is time consuming, and trial and error necessary.

***** Scope the project to the time available. Hone/narrow the question. Try to determine which might the most critical tests, or a subset of the data, and move on to those so you have plenty of time for interpretation. Don't get bogged down in technical details. Don't spend too much time transforming data to make it "normal."

* Form a question that leads to a conclusion, not an open-ended question that leads to exploratory research.

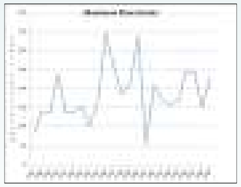


Figure 1: Extreme Flow Events

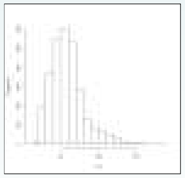


Figure 2: Histogram, of salamander Snout Vent Length across years. Is skewed only slightly to right.

* Number of asterisks next to a point indicate strength of student agreement.

The authors thank Mark Schulze, Forest Director, H.J. A. LTER, for help on this project. We also gratefully acknowledge participating students: Timothy Benedict, Jason Cornell, Eryn Farkas, Lola Flores, Lucy Gelderloos, Erin Hanlon, Evan Hayduk, Jennie Husby, David Kangiser, Heather Kowalewski, Ryan Kruse, Jason Lim, Evan Mangold, Zachary Maskin, Melissa Pico, Jesse Price, Matthew Ritter, Tim Rogers, Kari Schoenberg, Gregory Schultz, Allison Smith, Garrett Starks, Scott Stavely, Jerilyn Walley, Sarah Weber, and Marisa Whisman. Photo credits: Tim Benedict.

For more information, please contact PI Judy Cushing at judyc@evergreen.edu.

Metadata Commentary*

********* Metadata were comprehensive & clean.

******* We needed improved geolocation! Where were those study sites? Include maps with metadata! Make transect names consistent across publications.

**** We needed more information about methods. Why were certain protocols used? Why weren't all data measured across all plots?

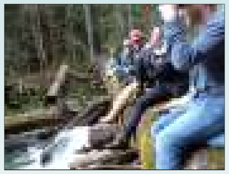
**** We needed better definitions and descriptions of (or universal) categorical data and variable names.

* Which tests were run and with what results?

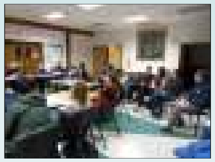
* Please provide links to supporting articles.

* Metadata were too discipline specific.

* Some metadata were misleading; our analysis showed different results than the metadata suggested.



	Cyano Lichen	Forage Lichen	Matrix Lichen
p-Value	0.046*	0.031*	0.001*
Age 1	0.95	0.67	0.67
Age 2	0.2	0.99	0.41
Age 3	0.03*	0.41	0.0002*
Age 4	0.01*	0.0055*	0.005*



Lichen fnc'l groups have higher biomas in older stands

Student Lessons Learned*

Students generally commented that analyzing a real data set and doing useful work for the HJA were valuable to them. Most difficult for them were:

***** working in R,

*** organizing the data,

*** creating a research question and refining hypotheses,

** using our own research question, but others' data,

** deciding which statistical test to use

* interpreting data transformations and statistical results,

* their lack of background in the field,

* extracting historical data from the literature, and

* knowing when to use MS Excel and when to use R.