Data Lab 1: Redwood Trees

ACMS 40950/60876 (Spring 2025)

Due Sunday, February 9, 2025 5:00 PM via GitHub

Background and Goals

Part 1: Problem Formulation and Data Collection. As an applied statistician/data scientist in today's data-rich world, we often collaborate (and if not, should be encouraged to collaborate) with people of domain expertise in order to answer scientific questions of interest outside of statistics using data. Before even tackling the analytical questions, this collaborative process starts with (1) carefully formulating the domain problem of interest, (2) identifying the relevance of this problem in the broader domain context, and (3) understanding how the data was collected well enough to ascertain how it can (or cannot) be used to help answer the scientific question. In order to ensure that our findings are meaningful and impactful in practice, our answers to these questions should heavily guide the downstream trajectory of our data analysis.

In the first phase of this lab, you will have the opportunity to grapple with these questions and the art of problem formulation. You will also be able to practice communicating your formulated problem of interest to a general (non-expert) audience by writing a short, but publication-quality introduction.

Part 2: Data Cleaning and Exploratory Data Analysis. Next, we will simulate the act of receiving data in a collaboration, which then jumpstarts the data cleaning and exploration stages of the data science life cycle. These are arguably *the most* important and often time-consuming tasks in the life of an applied statistician/data scientist. Yet, they are also the most overlooked. In my experience, good data cleaning/exploration can easily consume 90% of your time on a project. However, this time spent is often worth it as it can be the main determinant of success in applied collaborations.

In the second phase of this lab, you will gain experience uncovering (numerous) common data issues in real-world data and learning how to resolve these issues via data cleaning. After cleaning the data, you will then tie everything together by performing an exploratory data analysis to extract scientific insights, related to your domain problem under study.

Instructions

In this lab, you will be engaging with a real-world environmental science application from Tolle et al. [2005]. At a high-level, your objective in this lab will be to conduct your own exploratory data analysis of the dataset from Tolle et al. [2005] to study a particular scientific question (of your choice) and to write a concise, coherent, and reproducible report, summarizing your most interesting scientific findings.

Part 1: Problem Formulation and Data Collection

Please first take the time to *carefully* read Tolle et al. [2005] in its entirety, paying close attention to the domain context and the data source/collection. This is essential to your success in this data lab. In addition, as you are working through this data lab, you are free and encouraged to ask the instructor any questions about the substantive background as if the instructor were your domain expert collaborator. You may also want to skim some of the references in Tolle et al. [2005] to obtain a high-level overview of the substantive field.

After reading Tolle et al. [2005], identify a particular scientific problem or research angle about redwoods that you would like to further explore in this lab. In your report (e.g., your `.Rmd`, `.qmd`, or `.ipynb` file), write a concise and coherent *Introduction* section, which addresses each of the following questions below (in whatever order feels natural to you):

1. What is the big-picture scientific problem that you are interested in?
2. Why is this scientific problem interesting or relevant to researchers in the field and/or general members of society?
3. Are there particular challenges that make solving this problem difficult? Here, you may want to briefly discuss prior work and their limitations.
4. What is the specific aim or contribution of this present work, as it relates to the big-picture scientific problem?

Next, take a quick look at the original data from Tolle et al. [2005], which can be downloaded from Canvas. The main files of interest are `sonoma-data-*.csv` and `mote-location-data.txt`. Be sure to read *all* accompanying documentation.

After your *Introduction*, add a section on *Data Overview and Collection*, which briefly addresses each of the following questions (in whatever order feels natural to you):

5. How was the data collected or generated? [Note: a general reader with no knowledge of the original study should be able to understand how the data was collected at a high-level by reading only your data description. This description should also provide enough information to understand the remainder of your analysis. You can omit technical details that aren't necessary to understand your analysis.]
6. Briefly describe the variables in the data, including how these variables were measured/collected and why they are of interest. At a minimum, please describe each variable used in the second phase of this data lab, and make a concerted effort to justify why at least one of the variables or the set of variables as a whole are of interest.
7. Are there any special properties of the data/data collection that make it uniquely suited to answer your scientific question of interest? At a minimum, be sure to relate the data to your scientific problem.
8. Are there any limitations when using the data to answer the domain problem of interest? Be as transparent as possible so that conclusions made from this data are not misinterpreted down the road.

Data Cleaning and Exploratory Data Analysis

Your next task is to dig into the data and check its data quality. Do not take data consistency or correctness for granted. This data set is incredibly raw and contains some gross outliers, inconsistencies, and lots of missing values. Include a *Data Cleaning* section in your report, which contains the following:

9. Discuss *all* inconsistencies, problems, and oddities that you find with the data. These may include, but are not limited to: missing values, erroneous measurements, and outliers.

10. Bearing the data quality in mind, do your best to resolve these issues and appropriately clean the data.

    (a) Record in your report the steps you took to clean the data. These steps should be written in sufficient detail such that the reader can exactly reproduce your data cleaning procedure based upon your description.

    (b) Please also explain why you cleaned the data in that way when possible. Here, you may choose to bring in domain knowledge (e.g., from Tolle et al. [2005]) or common knowledge to support your data cleaning decisions. You may also find it helpful to include plots to help justify your data cleaning choices.

    - Note: You may want to read the *Outlier rejection* section in Tolle et al. [2005] carefully and critically, but do no blindly follow their data cleaning method.

Having cleaned the data, you can more easily explore and visualize the data in order to gain scientific insights related to your domain question of interest. In the *Exploratory Data Analysis* section, please include the following:

11. Perform an exploratory data analysis and identify 1-2 of your most interesting findings (one is required, the second is *extra credit*). For each of your interesting findings,

    (a) Produce a publication-quality visualization.

    - This is where I expect to see very polished graphics. Think carefully about use of color, labeling, shading, transparency, etc. This is your chance to do something innovative. If you are feeling bored or ambitious consider doing something dynamic or interactive. Publication-quality figures should also include an informative caption that explains how to interpret the figure and the main takeaway.

    (b) Provide sufficient detail regarding your analysis process such that a general reader could exactly reproduce your finding/figure.

    (c) Briefly discuss your finding in the context of the original domain problem of interest. Why is your finding scientifically interesting?

    Remarks:

    - A thorough and deep data exploration is necessary in order to arrive at *interesting* findings. This involves exploring/creating far more than two data visualizations, but without the need for publication-quality details.

- For your interesting findings, you may choose to show general patterns or anecdotal events. If using the entire dataset is challenging, try focusing on a subset of sensor nodes or a day's worth of data. Again, record in your report your process. Don't be afraid to try methods that are new to you and be critical of your own graphics.
- Finally, as you explore the data more deeply, you will most likely want to revisit parts of phase 1 of this lab and/or the data cleaning. This is completely normal and even a positive sign! You've just experienced the iterative nature of the data science life cycle first-hand.

Submission Details

Please push a folder named `lab1/` to your `dsip` GitHub repository **by 5:00 PM on Sunday, February 9, 2025**. I will run an *automated* script that pulls from each of your GitHub repositories promptly at 5:01 PM and attempts to reproduce your report.

The structure of your `lab1/` folder should follow the project structure discussed in class:

```
lab1/
 |__data/
 |__R/ (for R users)
 |__python/ (for python users)
 |__scripts/
 |__notebooks/
 |   |__lab1.qmd (or .Rmd or .ipynb)
 |   |__lab1.html (or .pdf or other rendered output)
 |__renv/ (for R users)
 |__renv.lock (for R users)
 |__environment.yml (for python users)
 |__conda-lock.yml (for python users)
```

The `R/` and/or `python/` folders should contain all functions (and only functions, no scripts) necessary to reproduce your report. The `data/` folder should contain the raw redwood data files, but **do not** push the `data/` folder to GitHub. In general, it is not good practice to store data on GitHub due to their restrictions on maximum file size (max: 100MB).

(Optional) You may include other files such as `README.md`, `LICENSE`, and `.gitignore`, which are considered good practice and very common in project workflows. You may also add other folders or rename the folders as you wish, but **do not** rename the `notebooks/` and `lab1.*` files since my automated script will be specifically looking for these file/folder names.

I will attempt to reproduce your report by clicking 'render' in RStudio or 'run all' in Python, so please be sure to include all necessary code in your repository. Keep in mind that my `data/` folder will only contain the raw data files that were initially provided to you.

A Note on Grading + Rubric

You will be graded on the content quality, reproducibility, overall narrative, and readability/grammar. Similar to writing a research manuscript, the overarching narrative and the readability/grammar are important components to good communication. While we do not expect you to be an expert on redwood trees a priori, we do expect you to learn a little about redwood trees through reading Tolle et al. [2005] and to incorporate some bits of this domain information throughout your report. Ideally, you should try to ground your discussions of your findings/analyses in the domain context, as if you were a true environmental scientist. A great report will also tell a story, where the writing flows from one section to the next and each plot has a reason for being included.

A detailed rubric can be found on Canvas.

Please recall the course policy regarding collaboration: Collaboration *of ideas* with the instructor and with classmates is encouraged throughout this course, with the following caveats:

- You must write up the final code and text by yourself.
- If you collaborate or use any resources other than course texts, you must explicitly acknowledge your collaborators (e.g., in writing at the end of your report) and cite the resources you used.

References

G. Tolle, J. Polastre, R. Szewczyk, D. Culler, N. Turner, K. Tu, S. Burgess, T. Dawson, P. Buonadonna, D. Gay, et al. A macroscope in the redwoods. In *Proceedings of the 3rd international conference on Embedded networked sensor systems*, pages 51–63, 2005.