Lab 1 - Redwood Data, Stat 215A, Fall 2021

Sahil Saxena

September 17, 2021

1 Introduction

Things to potentially include in your introduction:

- Describe the premise of your exploratory data analysis and put your analysis in the domain context
- Explain why studying this redwood data interesting and/or important
- What are the implications of better understanding this data?
- What is the purpose of your exploratory data analysis?
- Outline what you will be doing in the rest of the report/analysis

This report analyzes, critiques, and expands upon "A Macroscope in the Redwoods," a paper by Gilman Tolle et. al. which discusses data about single redwood tree in northern California. Specifically, the report explains the data collection from a multitude of sensors over a period of more than one month and presents its findings from this complex, multi-dimentional data.

*** Add more Here ***

2 Data

A team of scientists from University of California, Berkeley collected and analyzed data from a redwood tree in the Grove of Old Trees in Sonoma, California. The team built a wireless sensor network by installing sensor nodes, placed around the physical structure of the tree. They also used a local data logger to record readings from other sensor nodes.

The team chooses this data to measure by balancing the limitations of technology and the requests from local biologists. The team decided that this data would best give insights about the ecophysiology of coastal redwood forests. Although just studying one tree, this data helps biologists understand the spatial climate gradients around a large redwood tree and the temporal dynamics. For example, warm temperature fronts move down the tree over time and high humidity fronts move through the canopy over time. Some shortcomings of this data collection include only studying 1 tree (which prevents understanding variation over different tree types), not getting any direct solar radiation measurements (which forces bioligists to estimate the true sunlight), and lack of air pressure readings.

2.1 Data Collection

The data from each sensor was carefully routed via a mesh netowrk. Through this, the data was linked to a database running on a gateway. They also included a local data logging system in case of any network failure. They then ran simple SQL queries to select the relevant values.

The sensor readings were aggregated in 2 datasets: wireless sensor network and data logger. Data logger includes readings from 39 nodes placed on the "edge" (radially 1m from the tree) and 30 nodes placed on the "interior" (radially 0.1m from the tree). The wireless sensor network includes readings from 29 nodes on the "interior".

Each node had a battery and two sensor boards; one board captured radiation, the other captured temperature and humidity. The result was 4 value readings from each node: temperature, humidity, incident photosynthetically active radiation (PAR), and reflected PAR. By planting sensors at every 2 meters of height (between 15m and 70m) and at both 0.1m and 1m away from the tree, they ensured data which considers spatial variation. A majority of these sensors had to be placed on the tree's west side to gain protection from the tree's thicker foliage. Temporally, the data was recorded at 5 minute intervals from April 27 to June 10, 2004, giving a potenial total of 1.7 million data points. However, the logger data has only 301,056 observations and the network data has only 114,975 observations. Clearly, there are a lot of missing values and faulty data points.

2.2 Data Cleaning

- Discuss all inconsistencies, problems, oddities in the data (e.g., missing data, errors in data, outliers, etc)
- What steps did you take to clean the data, and why did you clean the data in that way?
- Record your preprocessing steps in a way such that if someone else were to reproduce your analysis, they could easily replicate and understand your preprocessing
- You may find it helpful to include relevant plots that help to explain the choices you made when cleaning the data
- Be transparent! This allows for others to read your work and make their own educated decisions on how best to preprocess the data.

The biggest problems in the dataset are missing values, nonsensical values, and an abundance of outliers, all of which make analysis impossible. The data contains the following meaningful variables:

Variable	Meaning	# of Missing (NA) Values
result_time	time of each measurement	-
epoch	ID of each measurement	-
nodeid	unique ID for each node	-
voltage	voltage of node	-
humidity	relative humidity (%)	12,532
humidity_temp	temperature reading (^{o}C)	12,532
hamatop	incident PAR	12,532
hamabot	reflective PAR	12,532

There are also variables called parent and depth, but the paper does not give a description about how to work with these. These 2 variables are removed from the data. Out of the 416,031 total measurements (including network and logger), there are 12,532 missing values for the climatic variables.

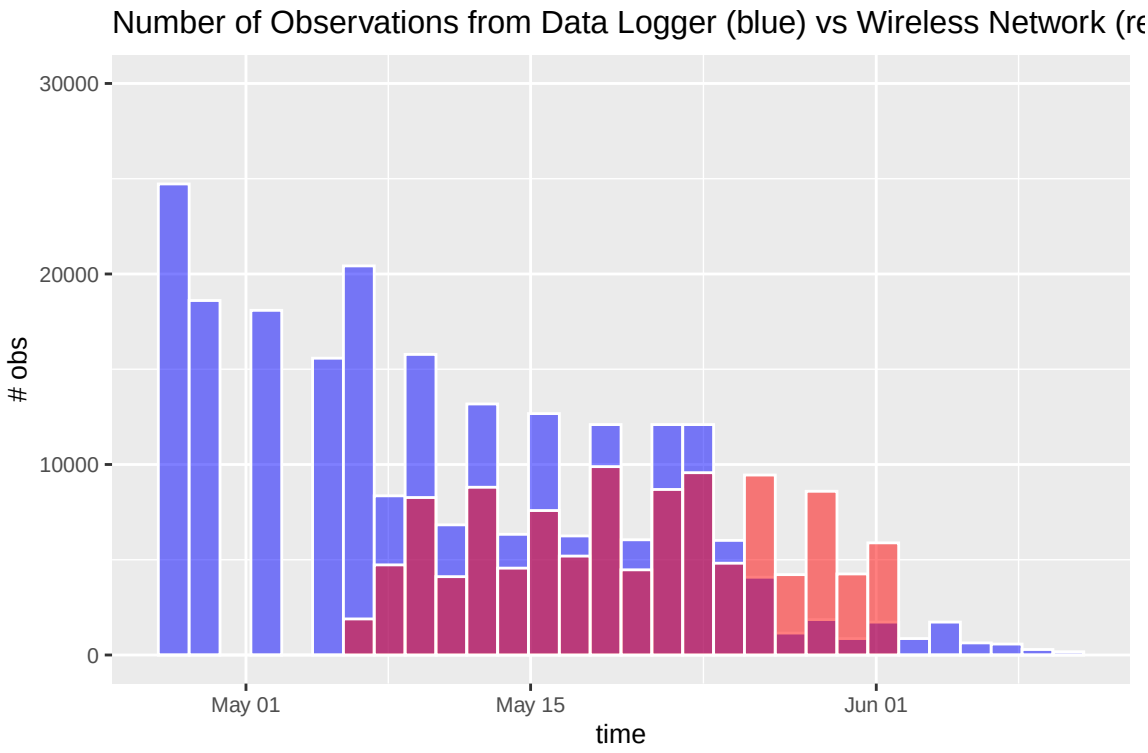

Figure 1: Histogram showing the number of observations for the net dataset (red) obverlaid onto a histogram showing the number of observations for the log dataset (blue) over time

It is also interesting to look at how well each of the nodes logged data. We can see that the data logger has more recorded observations, probably due to the fact that there are more nodes for logger. However, 28 of the nodes were common to both the wireless network and the data logger datasets. Of these 28 nodes, there are 10 nodes for which the data logger contained less than 1,000 observations and 8 nodes for which the network provided less than 1,000 observations, which means there is a lot of missing data here.

The number of overall observations made over time from both the data logger and the wireless network is presented in the figure above. At both tails of the time period, the data logger collected measurements while the wireless network did not. When collecting data, the number of observations from the wireless network was fairly constant, but definitely not constant for the data logs. This could be because if the wireless network is up and running, it will work consistently. However, the data logger was activated in case of any failures, which was most likely happening before the wireless network was set up.

After May 5th, the number of observations made daily approximately halves, and in the days following May 25th, the logger almost records 0 observations. According to Tolle et al., this is because data logs "filled up". Thus for the sake of insightful analysis, this report primarily examines the data logger dataset.

The variable humid_adj looked very similar to variable humidity, so humid_adj was also removed. There are also humidity values out of the 0 - 100 range, but this is nonsensical because this is a relative percentage. However, the humidity values above 100% visually seemed to fit with the data so those remained in the data, and could have happened due to a calibration error.

The logger dataset, the result_time variable had only one date (14:25:00, November 10th, 2004) repeated. Using the external file which had the correct epoch and result_time, the accurate times of measurement were put in.

The dataset has a large amount of missing values. This is because there are many node/time combinations not present in the dataset. Also there are 8,270 observations from 3 nodes in the log dataset with NA values

reported for each of the measurements taken, which had to be removed since more than 60% were NA. This is probably due to the node not functioning correctly for a large amount of time.

2.3 Data Exploration

- The main goal of this section is to give the reader a feel for what the data "looks like' ' at a basic level
- Think about plots that summarize the data, plots that convey some smaller findings which ultimately motivate your main findings
- A good report will tie everything together so that there is a reason for every figure in the story

2.4 Reality Check

- Do a reality check. What reality could you compare your cleaned data to?
- Clearly state your assumptions and explain why this reality check is useful.
- Does your cleaned data pass the reality check or are there issues? Discuss.

3 Graphical Critique

- Critique the plots in Figures 3 & 4 in Tolle et al.
- What questions did they try to answer? Did they answer them successfully?
- Did they raise any questions not addressed in the text?
- Would you change them at all?

4 Findings

- Present three interesting findings and produce a publication quality graphic for each along with a short caption of what each shows.
- Don't forget to appropriate label axes, titles
- Think carefully about use of color, labeling, shading, transparency, etc.
- Also interpret and provide an insightful discussion of what your figures show

4.1 First finding

Describe it and place a figure here

4.2 Second finding

Describe it and place a figure here

4.3 Third finding

Describe it and place a figure here

4.4 Stability Check

Take one of your findings and present a perturbed version. How does this affect your finding? Add a before and after plot here.

5 Discussion

- Did the data size restrict you in any way? Discuss some challenges that you faced as a result of the data size.
- Address the three realms: data / reality, algorithms / models, and future data / reality.

- Where do the parts of the lab fit into those three realms?
- Do you think there is a one-to-one correspondence of the data and reality?
- What about reality and data visualization?

6 Conclusion

- You should make attempts to connect your findings/analysis back to the domain problem in every section of this report, but here in the conclusion, you can reiterate your main points and provide overarching remarks on the redwood data as it relates to the domain problem

7 Academic honesty statement

Please address to Bin.

8 Bibliography