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

Sahil Saxena

September 18, 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

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.

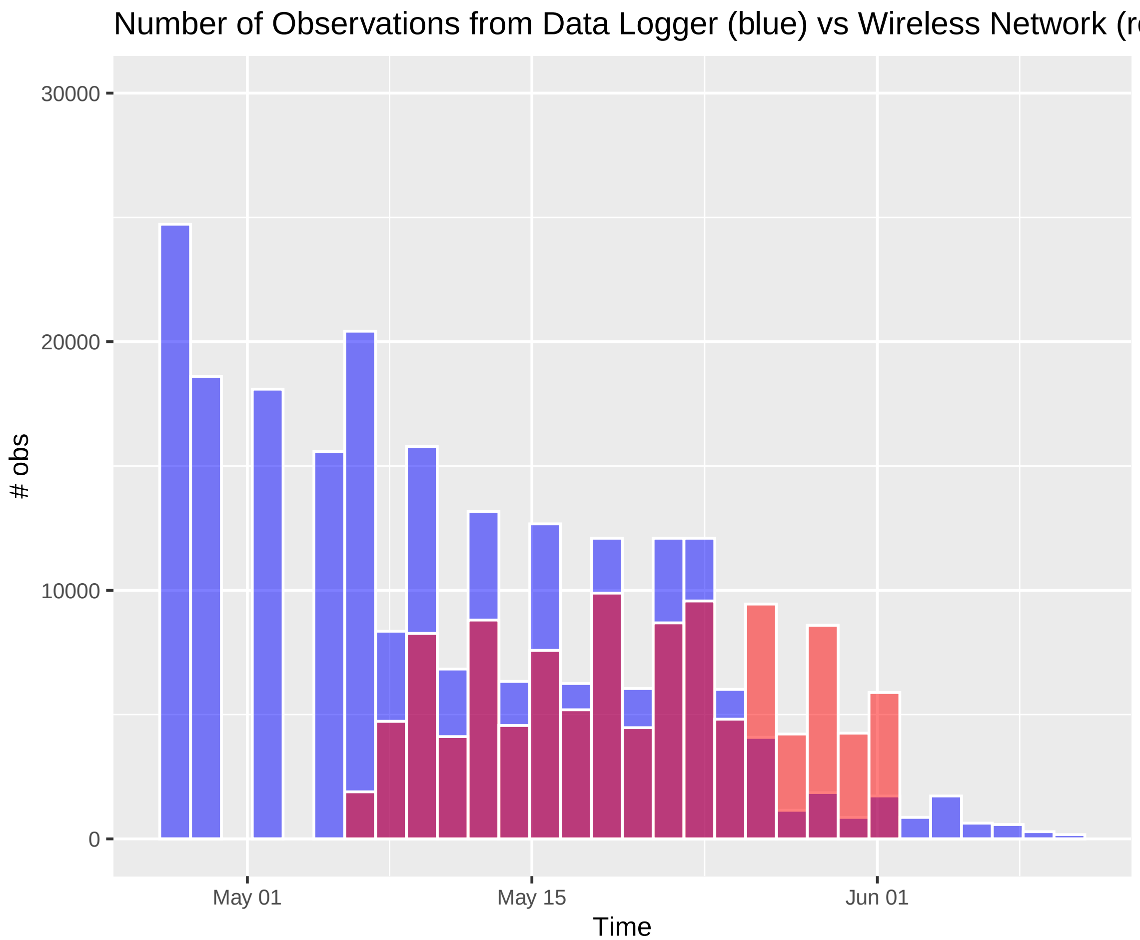

Figure 1: Histogram showing the number of observations for the net dataset (red) obverlaid onto a histogram showing the number of observations for the log dataset (blue) over time

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.

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 4 variables of significance are humidity, temperature, and both types of PAR. The scatter plots below show the dispersion of the observations over time. Each timestamp has a data point for all of its nodes.

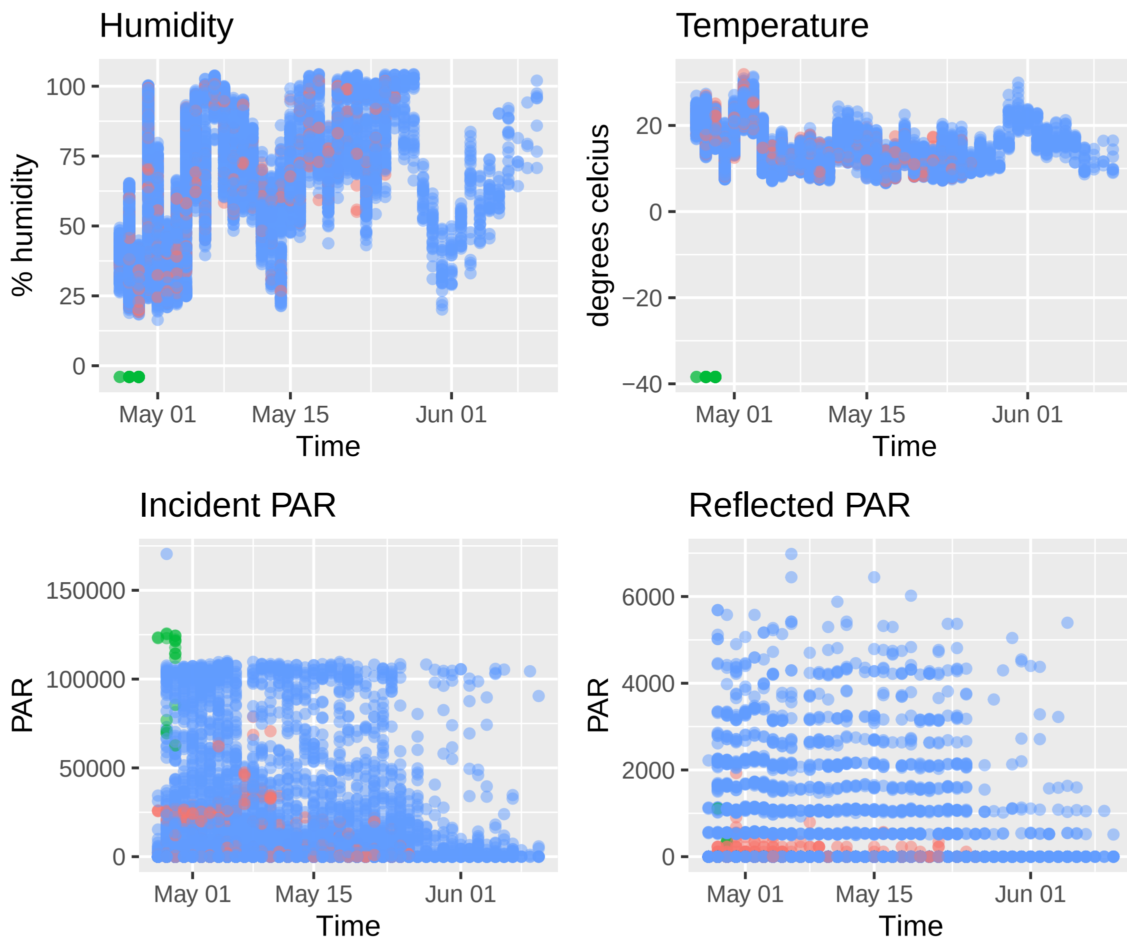

Figure 2: Scatterplot of each variable versus time in the data logger dataset. The color of each point corresponds to its voltage readings: red points have 0 voltage reading, green points are in which the node dies, and blue points have a valid voltage reading.

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 humidity values below zero correspond to a faulty node, which displayed strange voltage behaviour. These same nodes also recorded unreasonable temperatures incident PAR measurements.

This sheds some insight into what the data looks like, while accounting for the sensors' voltage malfunctioning. Temperature and humidity seem to follow some seasonal trend (here the season is over the course of a few days). Both readings of PAR seem to be somewhat uniform over time.

2.4 Reality Check

This cleaned data could be compared to other studies done measuring these 4 variables. Specifically, one could look at the same time period and compare these variables in other geographic locations. Also, it can be useful to look at the same month of the year, but in different years prior or after 2004.

By cleaning the data (removing values, ignoring certain time periods, etc.), underlying assumptions are that the remaining data is valid. This reality check can help verify whether these readings are really meaningful and valid. Just from looking at norhtern California weather historically, the values of temperature and humidity seem to make sense.

3 Graphical Critique

In Figures 3 & 4 of the paper, Tolle et. al. show many graphs. Figure 3 is about the 4 main variables and each is shown via a histogram, boxplots by day, and boxplots based on node height (one of these shows the difference from the mean). These graphs are intended to give the reader a visual feeling of the data, through the distribution of each variable and also through variability around the tree.

Putting the graphs next to each other helps raise questions about relationships between variables. Specifically, why does temperature and humidity have somewhat similar distributions, yet the PAR readings have a positive skew? Also, they raise the issue of humidity seemingly varying more as time goes on compared to temperature. Why does reflected PAR only have an abundance of observations for high node heights? The authors visually compare different variables to bring out these questions. Some can be answered visually and with background knowledge. For example, reflected PAR has variables at height because it measures reflected light, which is accessible away from the leaves in the forest. These questions also serve the purpose of a reality check. These questions also help to understand the “macroscope” and the general trends that occur in the forest.

Figure 4 highlights the variation in one day, which is important to understand as well. They remove focus on spatial features like height and ask the question about externalities, like sunlight, through the PAR time series. There are some unexpected results here, which is attributed to the uniqueness of this forest’s climate. Specifically, an inverse relationship is expected between humidity and temperature, but this redwood has large spikes in humidity witout large dips in temperature. This is how the authors move forward some of the larger findings; in this case, that the local climate has changed in strange ways.

I thought the graphs are informative and useful for setting up context. However, the two figures have some potential contradictions on height, which are not addressed in detail. Figure 3 shows the impact of height on variables, whereas Figure 4 downplays that variation and rather shows that temperature is fairly constant at all heights. Personally, I would choose to keep the factor of height involved in Figure 4 by using colors based based on height ranges. This would show that height still plays an effect and give a better overall summary of the true story.

4 Findings

4.1 Temperature and Humidity in times of Day

The figure below displays humidity versus temperature over four consecutive days and is supposed to highlight the time of day through coloring (dark points correspond to nighttime and light points correspond to daytime).

Note: I wasn’t able to debug this issue, so I left all points as the same color for now. Sometimes the graph would work correctly and sometimes it did not. Given more time, I would continue to troubleshoot this issue.

During daytime, the relationship between temperature and humidity is quite linear, whereas its not so linear during the night. During the night hours, every plot has 2 distinct temperature-humidity clusters, corresponding to early morning and late night. This is expected as they are separated by the daylight hours. Although the night still show an inverse relationship, the trend is not as steep nor linear as the day.

For example, the temperature decreases as humidity increases during the day on 05/01. However, at night, the temperature is fairly constant around 20, even with changes in humidity. At night, temperature is less responsive to changes in humidity. Also, there is a clear trend that the night is cooler and less humid than the daytime.

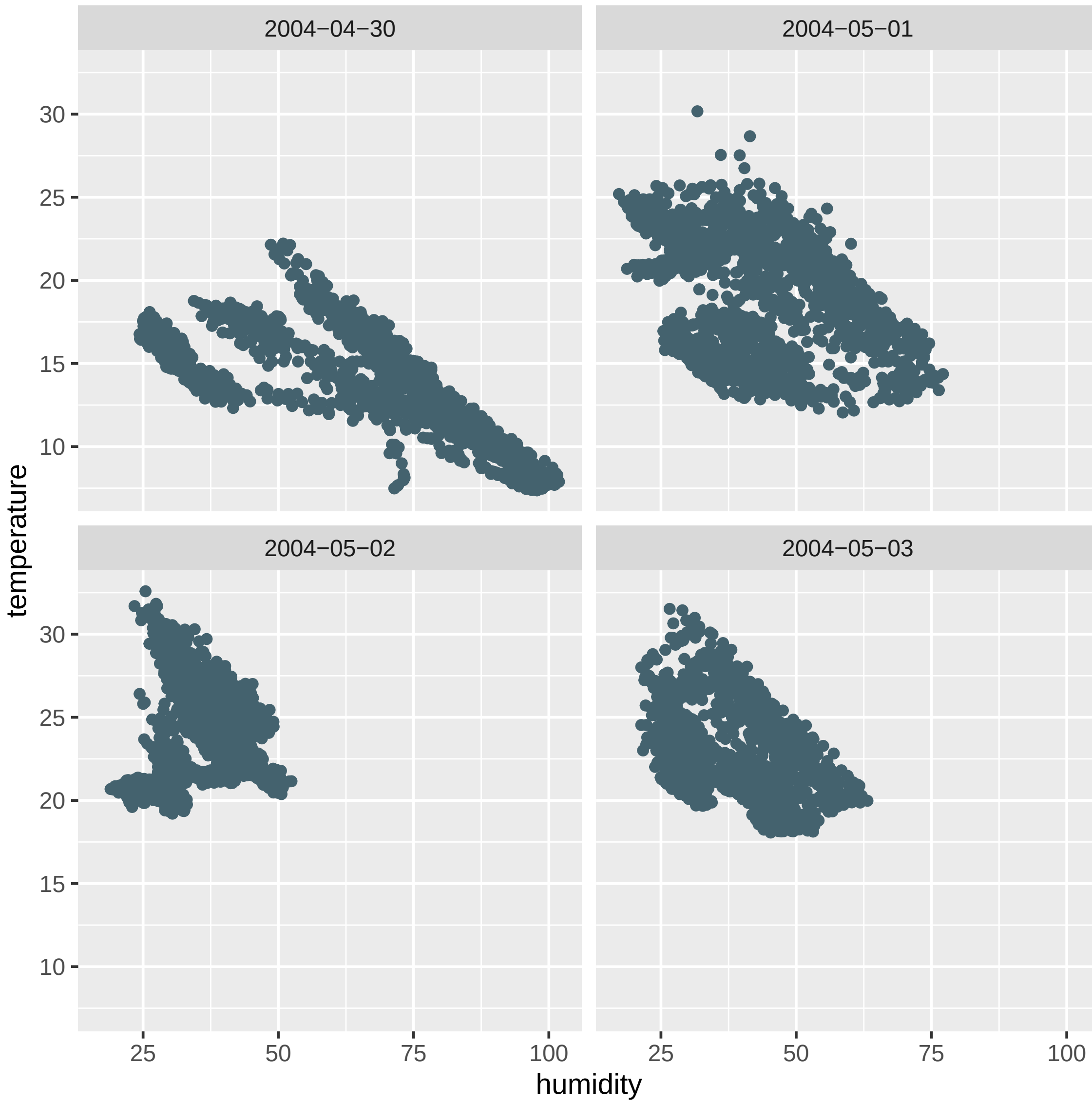

Figure 3: Temperature versus humidity for the interior tree on each of April 30th, May 1st, 2nd and 3rd.

4.2 Potential Discovery of Fog

This figure explores the edge and interior readings of temperature, humidity, and incident PAR on the dates of 4/28 and 5/4. 4/28 begins with a steep drop in temperature on edge and an increase in humidity. Meanwhile, the interior tree sees the opposite happen: a slight increase in temperature and a decrease in humidity. In the early morning, the temperature around the interior tree begins to decrease and then later rises. Looking at these variations, it is clear that the humidity will follow inverse changes.

While temperature and humidity are somewhat opposites, the incident PAR values over those two days show that edge sensors are picking up 0 values and interior sensors record regular PAR levels. One possible explanation for this pattern is that a cloud cover, or heavy fog moved in, which affected the edge sensors but couldn't get to the interior sensors. This might also explain the increase in humidity around the same time.

These graphs only show weather pattern on 2 days, but it seems to be the case that othese patterns match the behavior of the same variables on most other days as well.

Figure 4: Temperature, humidity and incident PAR (hamatop) over time on April 28th for which the two trees appear to experience very different environments and on May 4th when the two trees experience a very similar environment.

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