Lab 1 - Redwood Data, Stat 215A, Fall 2020

September 18, 2020

1 Introduction

As a native to Northern California, my childhood was occasionally punctuated by trips where I accompanied other children to be guided through the beautiful redwoods. The tallest and oldest trees of the world, they are home to their own individual ecosystems and microclimates that vary richly up and down the tree. And yet, between logging of the valuable wood, as well as California wildfires powered by climate change, the details are rubbed away.

My belief is that if we are going to allow something to decay or fade, we must do it consciously; that is, knowingly of what we are giving up. Hence the purpose of this report is to explore the climatic data of a redwood tree from a visual, easily-digested perspective, and to see pictures of a redwood tree's microclimate that will not be too quickly unseen. It is too sad to have things vanish, with us not even knowing what we have left behind as we roll and crash our way through history. As part of this, it is also important to understand the specific challenges of the data collection and retrieval that produces these pictures in the first place.

Hence in this report I will analyse the dataset referenced in Tolle et. al [1] of climatic data gathered from redwood trees in Sonoma, CA. I will describe the data, and issues in its preparation and organization; then, having fixed the issues, I will explore it from a birds-eye view, and verify its consistency. Then I will critique the plots of [1], that answers questions on the topic of this report; finally I will analyse the data in a new way and provide understanding of the redwood tree's microclimate.

2 Data

2.1 Data Collection

(I planned to discuss the data collection process from Tolle et al for PAR incident, but did not have time.)

2.2 Data Cleaning

The data as given had numerous issues. Here are some of the inconsistences uncovered in this analysis:

1. Two obvious outliers with extreme readings for most of the climatic sensors, namely the ones with humidity readings in the negative thousands.
2. Unreliable voltage readings, in the two hundreds of volts.
3. The PAR sensors were in unknown units.
4. There were many more nodes in the dataset than mentioned in [1].
5. Unreliable result_time column—some had values in november.
6. Many duplicates of epoch/node_id pairs, which had identical climatic measurements but different voltage and result_times.
7. Some duplicates of epoch/node_id pairs existed, but which had DIFFERENT climatic measurements.
8. Unreliable sensors that no longer measured correctly, and identifying them.
9. No documentation on what the columns parent and depth are, and some strange nodes that don't appear in the mote-location-data.txt.

Here is how they were resolved:

1. These were simply removed.

2. All of these voltage readings were from the data sent through the network, and it turns out that many (but not all) of them had duplicates in the logs (matching by node_id/epoch). It was also discovered that the strange voltage readings had a linear relationship with the voltage of their duplicates in the logs. A linear regression was run and the correct voltage was inferred and imputed by the linear relationship.
3. It was inferred that they were in lux, as are most photodiodes, and then to convert from lux to einstein units
4. There were actually TWO trees in the dataset, only one of which was analyzed in [1]. This was determined by comparing the heights of nodes to the ones used in the paper and closely reading mote-location-data.txt. Throughout the analysis I restrict myself to the one tree used in [1].
5. Even though [1] removed these by a voltage condition, I found this to be insufficient, because many low voltages were valid data points (as can be shown through tables) and many high voltages were actually good datapoints. The errant nodes were actually removed by making line plots and observing by hand when the nodes appeared to become erratic, and removing those data beyond when they appeared to stop tracking the other values. (See the below figure of line plots.)

(I did not have time to complete this section.)

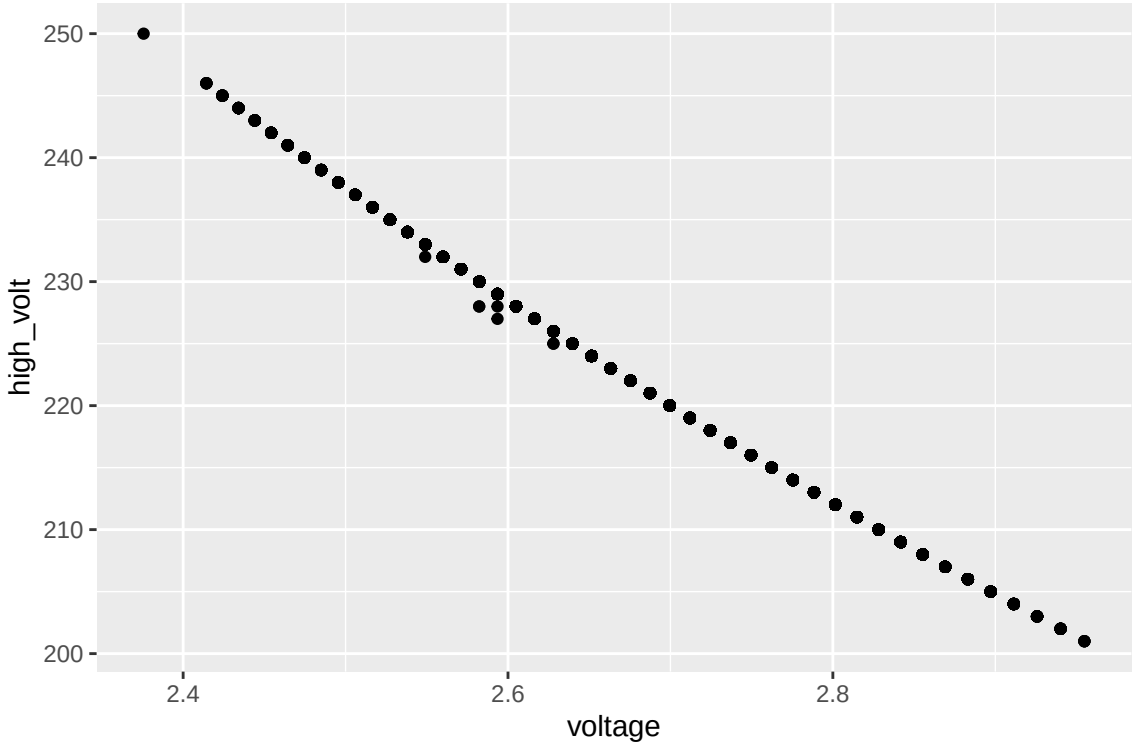

Figure 1: The linear relationship between the high voltages and the normal voltages.

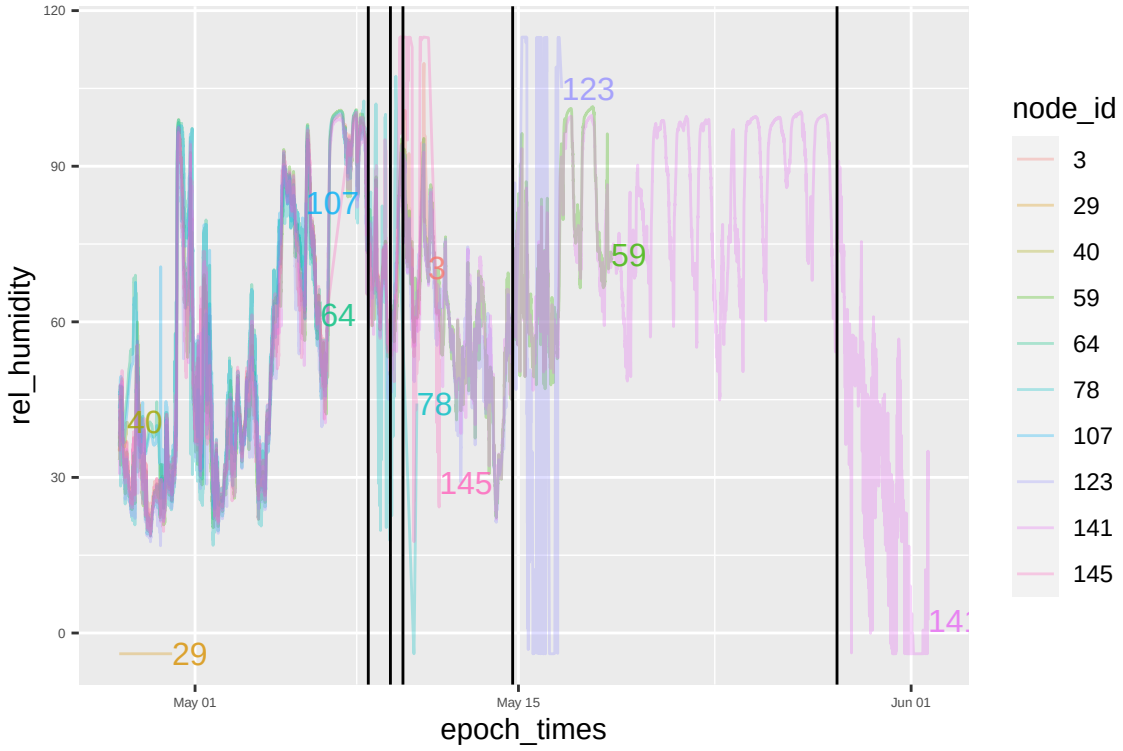

Figure 2: Plots of the nodes that went haywire, and the timepoints after which I scrubbed their data.

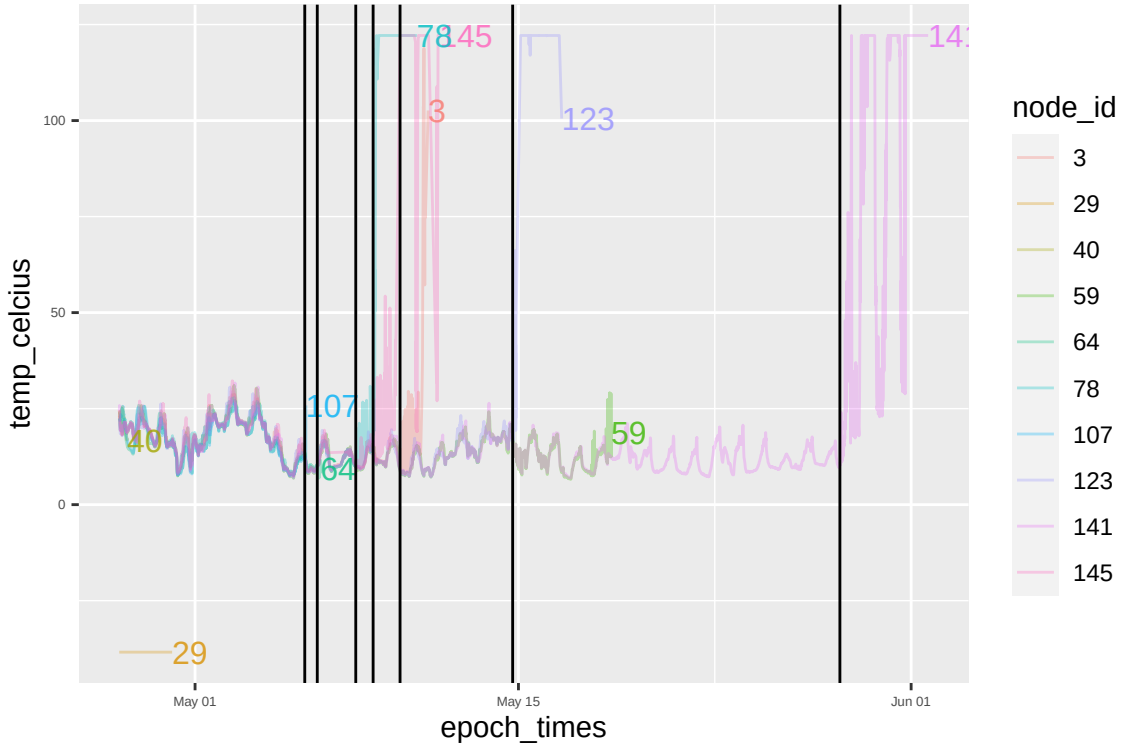

Figure 3: Plots of the nodes that went haywire, and the timepoints after which I scrubbed their data.

2.3 Data Exploration

Per the previous discussion, the data was cleaned and all referenced issues were fixed.

Visual summaries of the data will now be presented. As previously noted, the data was collected over 44 days from sensors attached to 33 nodes up and down the height of the 65m redwood tree, but only one node collected all 44 days of data; the others all gave out at earlier times.

Notwithstanding, we provide visual summaries of the six-and-a-half weeks of data collection. First we present plots of temperature and humidity, which are highly correlated environmental variables. One notes the large-scale changing weather patterns observed around the tree.

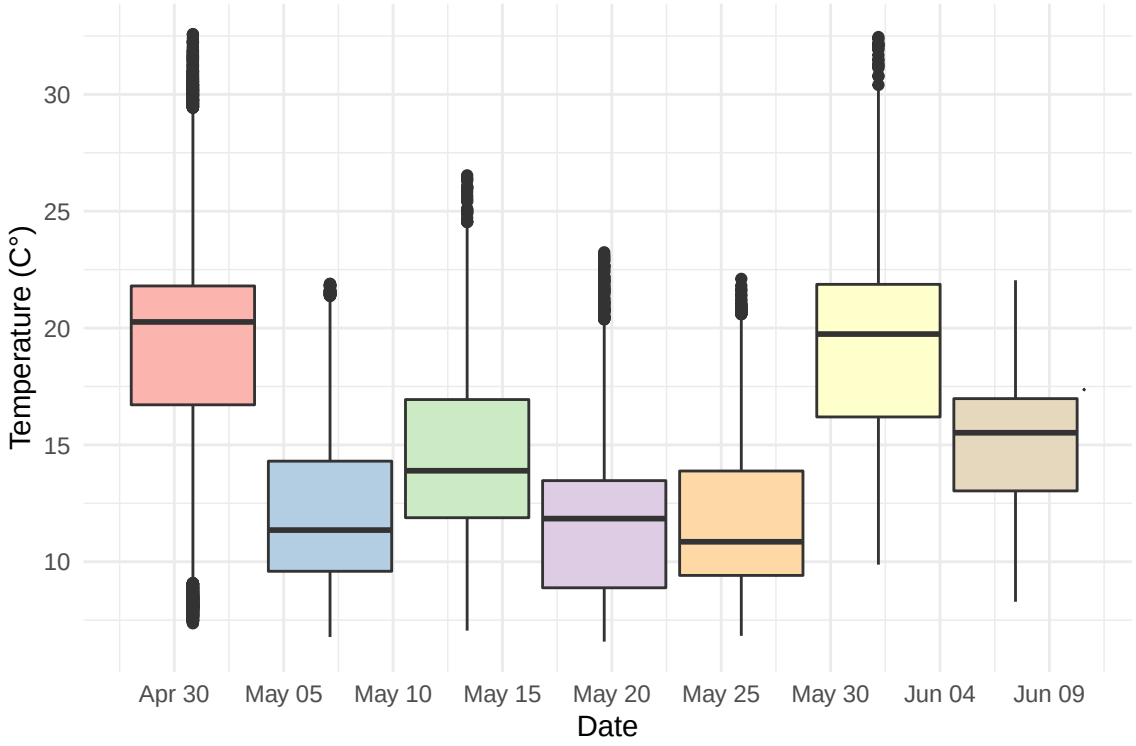

Figure 4: Temperature over the data collection period. Each box represents about six days.

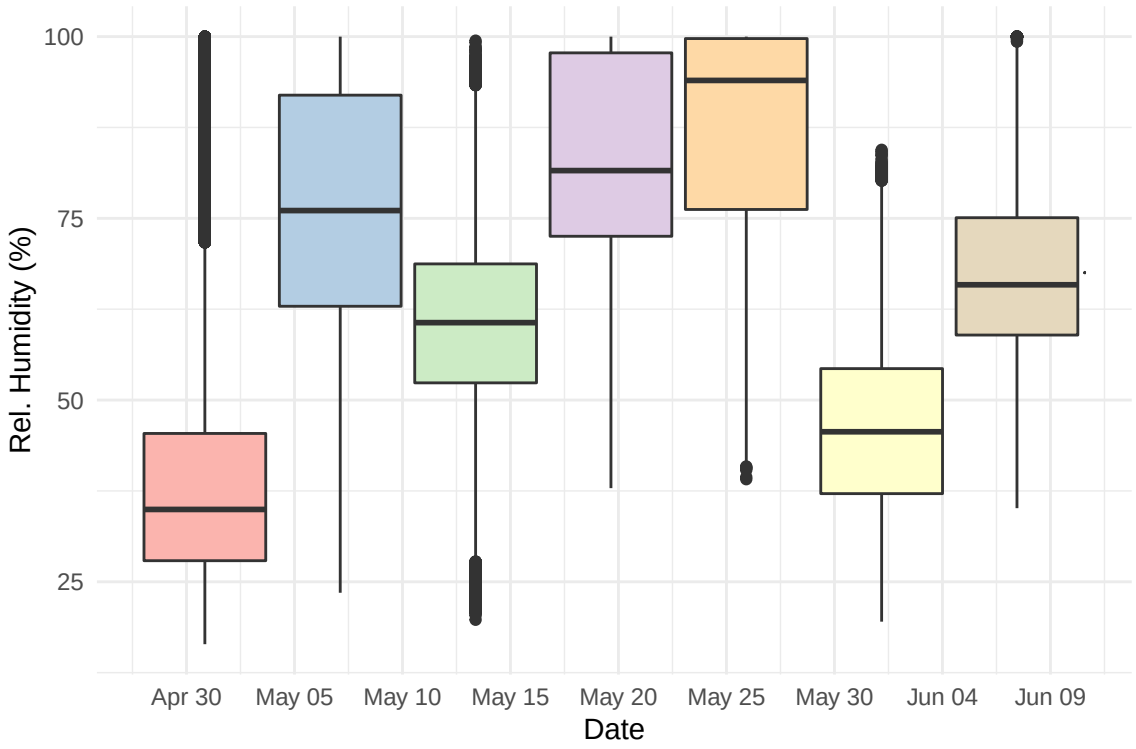

Figure 5: Relative humidity as a percentage.

(I also planned to display ridgeplots of the incident PAR over two days, but did not have time.)

2.4 Reality Check

The cleaned data seems reasonable—the humidity and temperature move in opposite directions over time, that is colder temperatures correspond to higher relative humidity. This makes sense, because for the same amount of water in the air, at colder temperatures the maximum water the air can hold is less, so that the relative humidity is higher.

3 Graphical Critique

4 Findings

4.1 First finding: Light through leaves

Here we display a heatmap representing the incident PAR in einstein units at every time point and tree height. Brighter colors represent more incident light and in fact ~2000 represents direct sunlight.

We see that the nodes must be positioned such that as the Sun takes similar paths across the sky each day, similar patterns of light emerge, for instance that only at sunset do the lower nodes get hit by the sun.

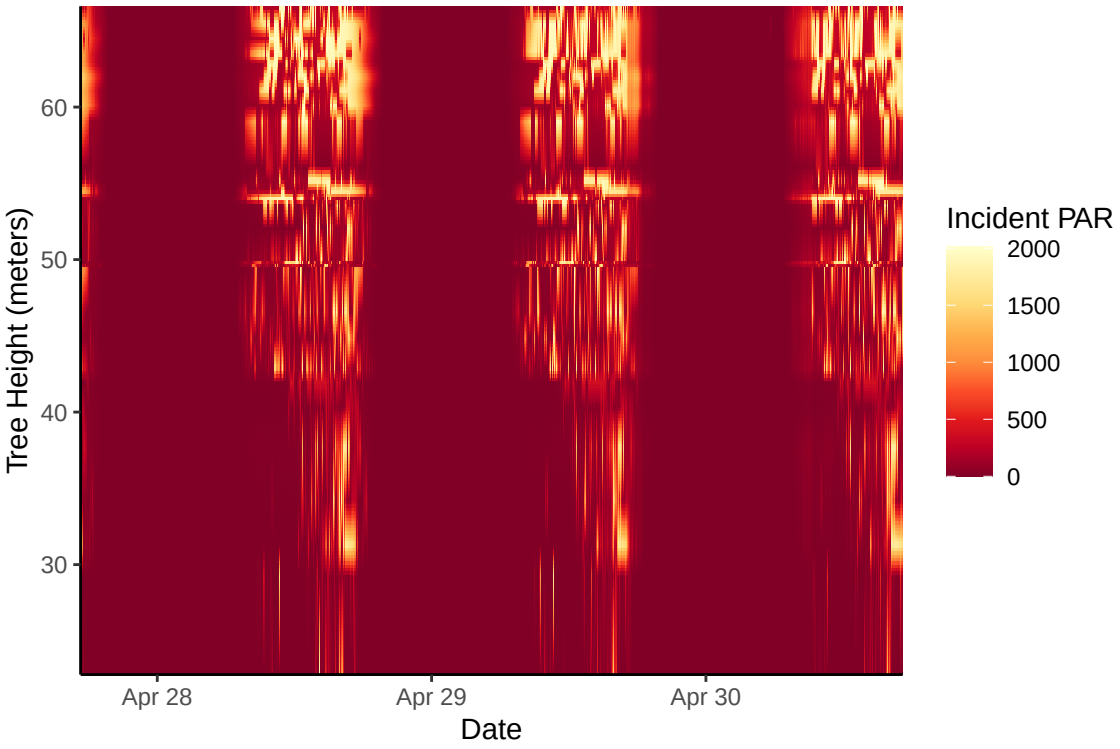

Figure 6: At every time point, the interpolated PAR

4.2 Second finding: Interaction between sunlight and temperature

(I was planning on looking at the delay between incident PAR and temperature warming up, as a way to analyze the normalizing effect the tree had on its microclimate, but did not have time.)

4.3 Third finding: Humidity vs temperature throughout a day

(I was planning on plotting a series of ridge plots but did not have time.)

4.4 Stability Check

(I was planning on rerunning the analysis, but this time with removing data that was above 3 volts, but did not have time.)

5 Discussion

6 Conclusion

Off topic...

I spent a lot of time data cleaning and could not organize enough to snap out of it and create the rest of the report. I actually think I cleaned the dataset too much because I cannot think of anything else to clean, and did not have time to generate plots for findings and write the text of the report.

My grade may not quite be proportional to the amount of sleep I lost in my efforts, but regardless I learned a lot from this experience. I was a little too perfectionist and did not give myself the appropriate time to *be* that perfectionist.

I'll get them next time!

7 Academic honesty statement

I believe in academic honesty and integrity. I believe that it is essential that what turned in reflects our honest work, because the same attitude ensures the integrity of science.

8 Bibliography

[1] Tolle, Gilman, et al. "A Macroscope in the Redwoods." Proceedings of the 3rd International Conference on Embedded Networked Sensor Systems - SenSys '05, 2005, doi:10.1145/1098918.1098925.