Lab 1

9/16/2021

Section 1. Introduction

In this report, I explore and analyze data on the microclimate surrounding a redwood tree. This dataset, collected by a wireless sensor network, is of interest because it enables us to peer into the day-to-day life of a redwood tree. My objective is to obtain specific insights on how the microclimate varies in time and space, which hopefully can be generalized to future ecological studies.

In my exploratory data analysis, I strive to familiarize myself with the major trends at play in the numeric, categorical, and temporal variables recorded. I also strive to distinguish between these trends and the random noise that we need to clean from our data. In this sense, data cleaning and EDA are inextricably linked: Through exploring our data, we are able to discern what components need to be extricated.

Section 2. Data

Our data comes from the paper *A Macroscope in the Redwoods* (Tolle et al., 2005). The climate conditions that the authors observed are temperature, humidity, incident light, and reflected light. To track the temporality of these conditions, they recorded time and epoch. And to take spatial dimensions into account, they observed each mote's height, orientation, and distance to the trunk.

Section 2.1 Data Collection

The data was collected from a wireless network of sensors placed at varying heights on a redwood tree. These sensors were placed roughly every 2 meters up the trunk of the tree, spanning from roughly 15 to 70 meters. They took measurements every 5 minutes for 44 days, from late April to early June.

The sensors were connected in a graphical structure by means of a routing tree. Every time a sensor recorded a new set of measurements, it would transmit these measurements through the routing tree to the root node. This node, a Stargate gateway, would then transmit the data over a GPRS cellular modem to some external database.

The wireless network, however, only recorded about 40% of the measurements. Because of this, Tolle et al. also backed up data to a logger at the base of the tree. While this logger recorded around three times as much data as the network, it ran out of memory for most nodes on May 26. These two datasets contain the environmental and temporal variables. The metric we use for light is photosynthetically active radiation, or PAR. For humidity, we

use relative humidity, expressed as a percentage. This measures the amount of water vapor in the air relative to the quantity that can exist at the current temperature. These datasets also contain voltage, which according to Tolle et al. ought to be between 2.4 and 3.0 volts.

In addition to the network and logger datasets, there is a dataset containing information on each sensor. This contains spatial data like height, orientation, and distance from the trunk. It also has a variable, "Tree," which can be either interior or edge. The meaning of this is not clear; we discuss it in Finding 3.

Section 2.2 Data Cleaning

Section 2.2.1 Time

Motes are supposed to take a measurement once every 5 minutes. However, the time readings are imprecise in both datasets. In the logger, every measurement is said to have happened in November, an obvious error. And in the network dataset, they are taken shortly or after the 5 minute mark.

To correct this, I rounded times to the nearest five minutes. In rounding, we are assuming that 1) the epoch readings are accurate, and 2) analysis of temporal data won't be perturbed by changing the time dimension by small amounts of time.

The first step is to convert the time column to standardized format with the lubridate package. Then, I reset each measurement's time with the following formula:

$$CurTime = BaseTime + Interval * (CurEpoch - BaseEpoch)$$

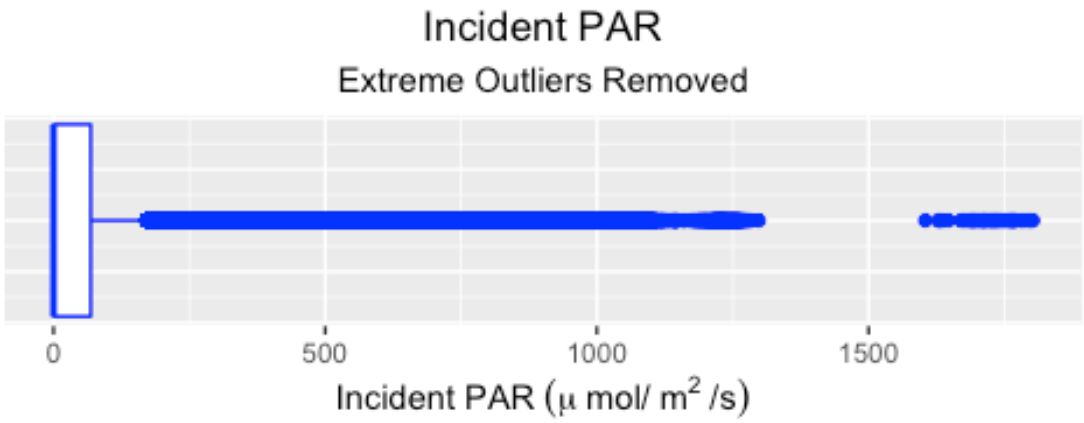
where the interval between epochs is 5 minutes.

Section 2.2.2 Duplicates

Both datasets contain thousands of duplicates. Identical rows can easily be removed in R using the unique() function.

Section 2.2.3 Light

PAR, our measurement for light, required extensive cleaning. A quick search revealed that our measurements exceeded the paper's by a factor of 100, so I normalized them as such. Then, after removing extreme outliers from incident PAR, I visualized the spread with the boxplot. As shown below, there exists a group of outliers above 1,500 $\mu mol/m^2/s$. This abrupt jump in values suggests that they are unwanted outliers, so I removed them. Reflected PAR has two extreme outliers above 1,000 $\mu mol/m^2/s$, both of which I removed. The remaining data was more evenly distributed, so I did no further cleaning.



Section 2.2.4 Voltage

The network and logger datasets have very different voltages. Experimentation in the paper demonstrates that the voltages are supposed to be between 2.4 and 3.0 Volts. To get the network voltages within this range, I normalized them by a factor of 100. Per the paper's suggestion, I removed all voltages not between 2.4 and 3.0 Volts. This was a judgement call that errs on the side of cautious. Voltages outside the standard range mean that either the sensor is malfunctioning or its battery is dying – both of which suggest that its performance is unstable.

As it turned out, almost 99% of the nodes in the network dataset have voltage outside of this interval. So in the interests of having plentiful data, I kept a second version with voltages between 2.1 and 3.3 Volts.

Section 2.2.5 Temperature

The temperature values are full of unreasonable outliers, from -40˚ to 600˚C. A boxplot shows outliers below 5˚ and above 35˚, so I removed the points outside this range. Most temperatures come from a narrower range of values; the inter-quartile range is between 10˚ and 18˚.

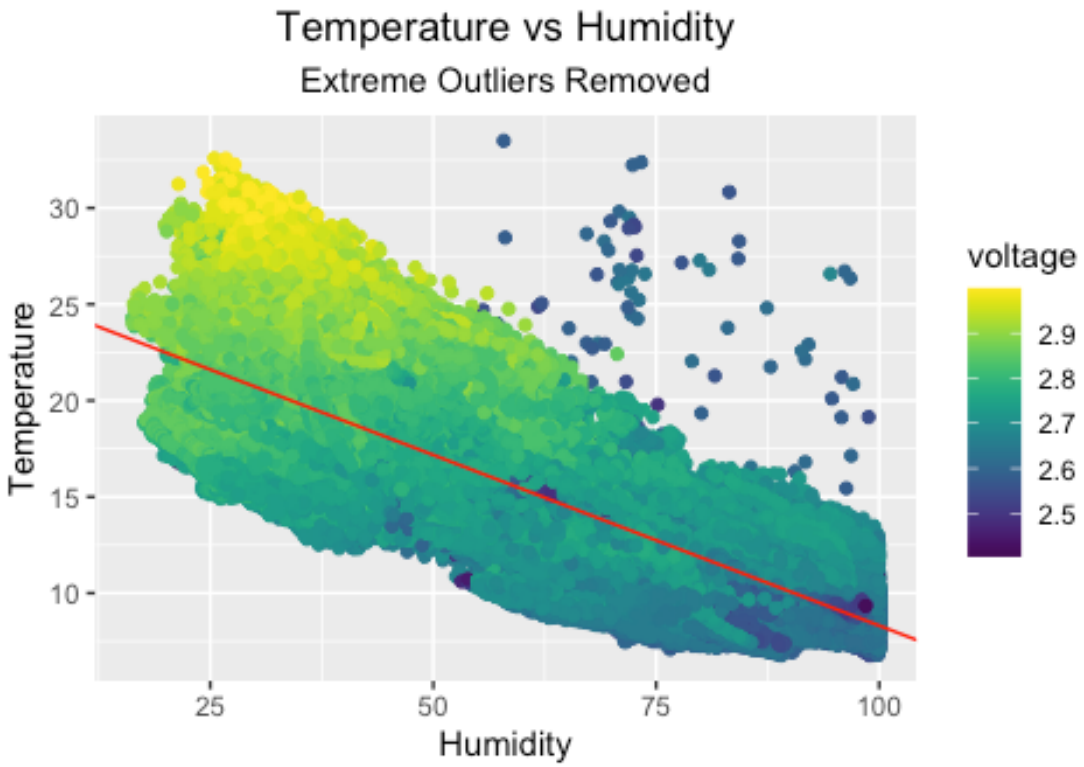
Section 2.2.6 Humidity

Relative humidity values represent percentages. Because of this, I remove all measurements whose humidity values are below 0 or above 100.

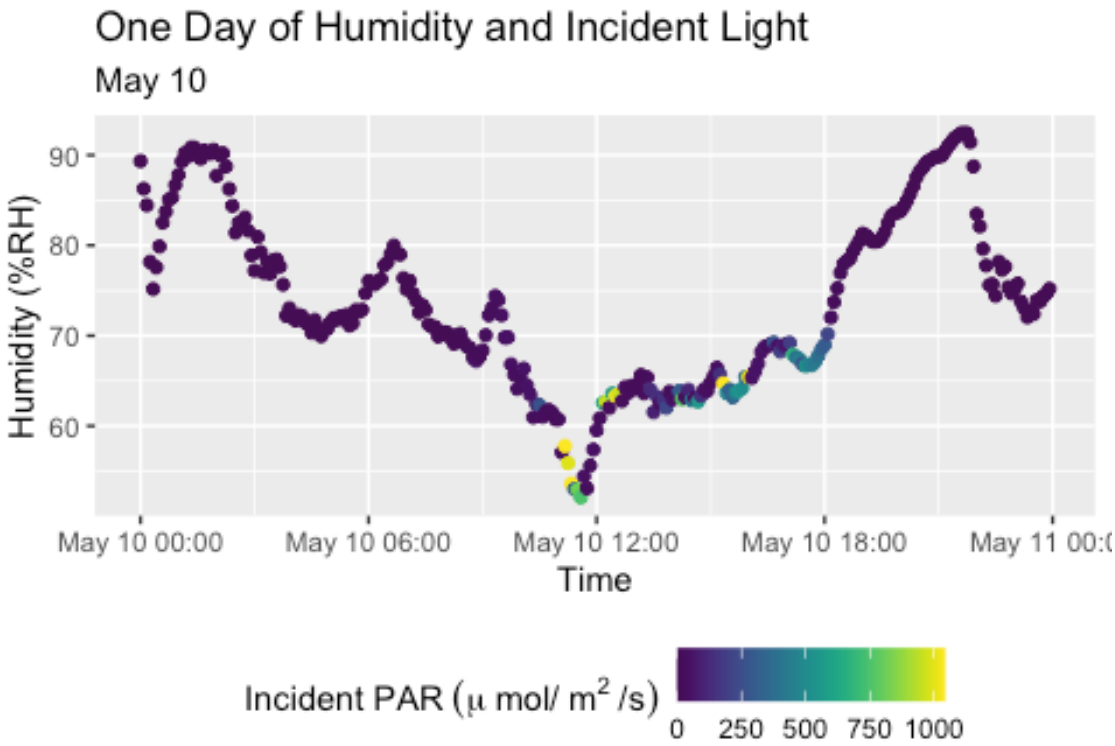
Section 2.3 Data Exploration

The first relationship we'll explore in our data is that between temperature and relative humidity. These variables share an inverse relationship. Usually, the amount of water in the air doesn't change dramatically over short periods of time. So when the temperature rises, the air is capable of holding a larger amount of water, thus leading to a decrease in relative humidity.

In the graph of cleaned temperature versus humidity, this relationship is generally present. The regression line, shown in red, tells us that for every 1˚ C decline in temperature, relative humidity increases on average by about 5%. However, above the blob highlighting this relationship are a couple dozen outliers. While there are not enough to dramatically change the regression line, they are clearly outliers that need to be cleaned. (I do so after showing this plot.) Note that higher temperatures are corellated with higher voltages.

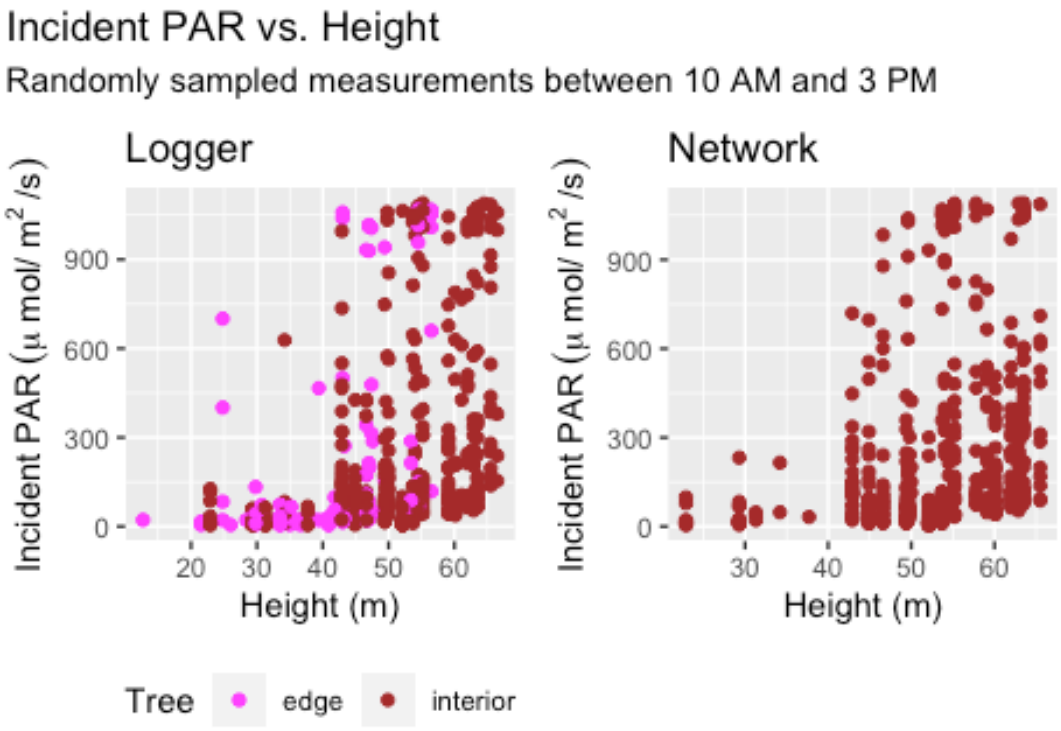


Next, I plot humidity and incident PAR over a 24-hour period. Incident PAR increases from dawn until midday, as the sun gets brighter and brighter; the sun brings warmth, which causes the relative humidity to fall dramatically. At noon, humidity sharply changes course, climbing back up to a peak of over 90% around 10 PM. It then drops steeply to a new baseline. Due to the time, this final decline probably cannot be attributed exclusively to temperature. Rather, it must be caused changes in the amount of water in the air. Knowing the Bay Area, the likely culprit is fog!



The side-by-side scatterplots below highlight the relationship between Incident PAR and height. As we might imagine, higher sensors tend to reflect more light, since they aren't as obstructed by leaves. This is not an absolute rule: Plenty of higher sensors reflect little or no light. Another interesting observation is that this trend only provides information up until about 50 meters.

The plots also highlight a key difference between the network and logger datasets. We can see that the logger plot contains nodes from both the interior and edge trees, whereas the network plot just contains points from the interior tree. Indeed, only 3 of our roughly 80,000 measurements in the cleaned network dataset are in the edge tree. So in general, the network data records one tree, and the logger records two.



Section 2.4 Reality Check

Many of the temperature values I cleaned were clearly detached from reality. Temperatures above 60˚ C, for example, would be unheard of anywhere on Earth. Having lived most of my life an hour's drive from the redwood grove, I can give a more in-depth reality check. My lower bound of 5˚ C is perfectly reasonable, as I've never experienced springtime temperatures that cold in the Bay Area. Nor have I ever experienced springtime temperatures above 35˚ C, the upper bound I chose.

Moreover, a quick online search found that as of 2010 the hottest temperature ever recorded in Healdsburg was 38˚ C. Healdsburg is decently far inland from the redwood grove, meaning its temperatures are going to be hotter. Given this experience and knowledge, I think it was reasonable to restrict temperatures to within 5˚ and 35˚. Such a restriction was neither too conservative nor too aggressive.

Section 3. Graphical Critique

Figure 3 displays all four of our variables – temperature, humidity, incident PAR, and reflected PAR – in a one- and two-dimensional setting. The one-dimensional visualizations (Figure 3a) are histograms that show us the basic distribution of the values. They reveal, for example, that incident and reflected PAR are skewed right. Figures 3b and 3c plot each measurement against time and height, respectively. This gives a good big-picture overview of the data, although plots like humidity versus height can be so jam-packed with data that it's difficult to discern the major patterns. Lastly, the authors wondered if readings'

deviations from the mean can be explained by height. Figure 3d answers in the affirmative for incident and reflective PAR.

Figure 4 displays all four environmental variables with respect to both time and height.. For example, they successfully found an inverse relationship between relative humidity and height for a fixed moment in time. This relationship is non-linear: The decreases in humidity get progressively larger as height increases.

Another question they sought to answer was how and why incident and reflected PAR differ. They averaged their measurements over time with a smoothing line, which effectively conveys the overall patterns of the light. They observed that incident PAR adheres to a roughly normal distribution during the day. This makes intuitive sense, as daylight is lowest at night, transitions at dawn and dusk, and peaks in the middle of the day. Reflective PAR, in contrast, rose slowly from sunrise until the late afternoon, then sharply dropped off as night fell. Tolle et al. reasoned that this was due to the fact that the tree's sensors were on the west side. They got increasingly more direct light over the course of the day, since the sun crosses the sky from east to west.

While the plots with regards to height are interesting, I personally would have flipped the axes. Of course, it is visually appealing to see height on the vertical Y axis. However, it doesn't make logical sense, as the Y variable is supposed to be dependent on the X variable. In this case, height causes environmental factors like humidity to change, not the other way around.

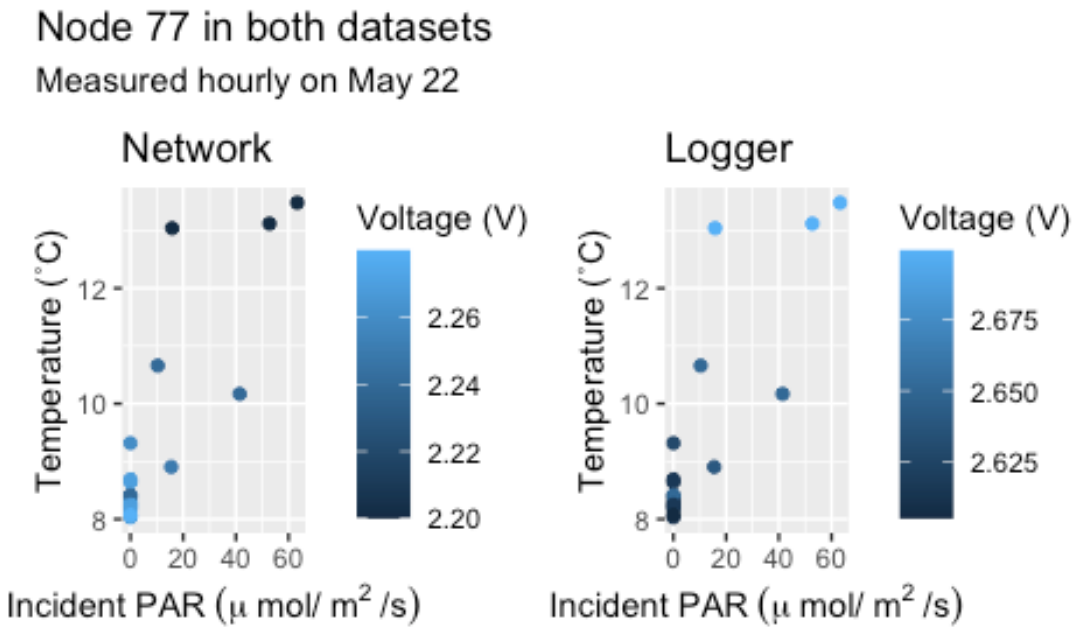
I also would have removed the outliers from the plot of Incident PAR vs height. In Section 2.2.3, I discussed my decision to remove values of Incident PAR above 1,500 $\mu mol/m^2/s$, which I concluded were erroneous outliers. In Tolle et al.'s plot, we can see how far these four values are from the rest, and how heavily they affect the linear regression. A line fitted without them would clearly have much lower residuals.

4. Findings

4.1 Finding 1

The graphic below reveals that there is significant overlap between the network and logger datasets. Here, we display side-by-side plots of hourly temperature and light measurements on May 22 for a node that appears in both datasets. We can see that the two plots show the same data.

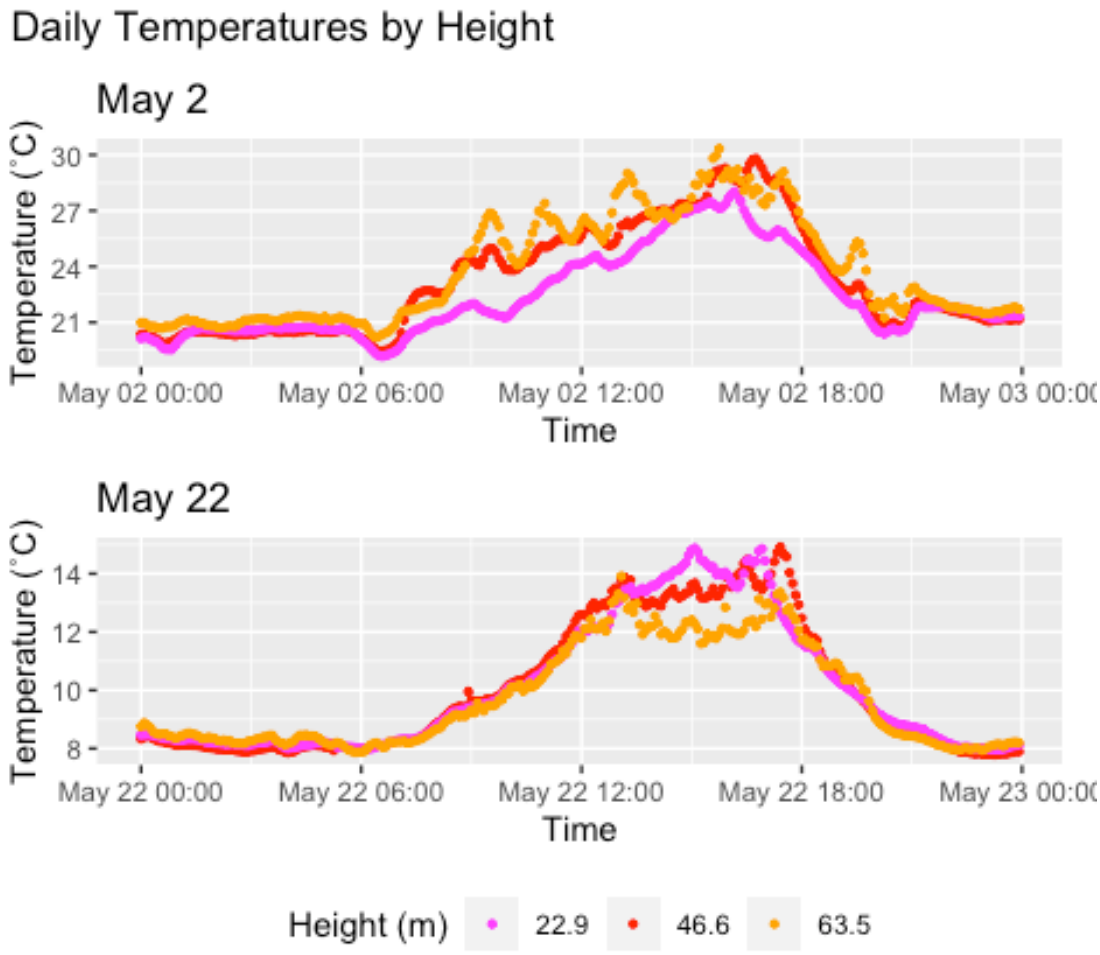
Surprisingly, the voltage is different across the two datasets. The recordings come from the same node, so the voltages should be identical. What's more, all the network voltages fall in the "danger zone" of below 2.4 Volts. This suggests that the wireless network has a problem with recording voltages, not with the voltages themselves.



4.2 Finding 2

These next plots explore the effect that height has on temperature throughout the day. May 2 was a hot day, and for most of it, the higher sensors recorded hotter temperatures. This makes intuitive sense, as the lower a node is, the less light is able to permeate through the forest. Differences could be substantial, especially as the day was warming up. At most, the high and low sensors differed by 5˚C. Higher sensors are also more volatile to changes in temperature. Clouds going over the sun, for example, wouldn't have as much effect on an already-shady area.

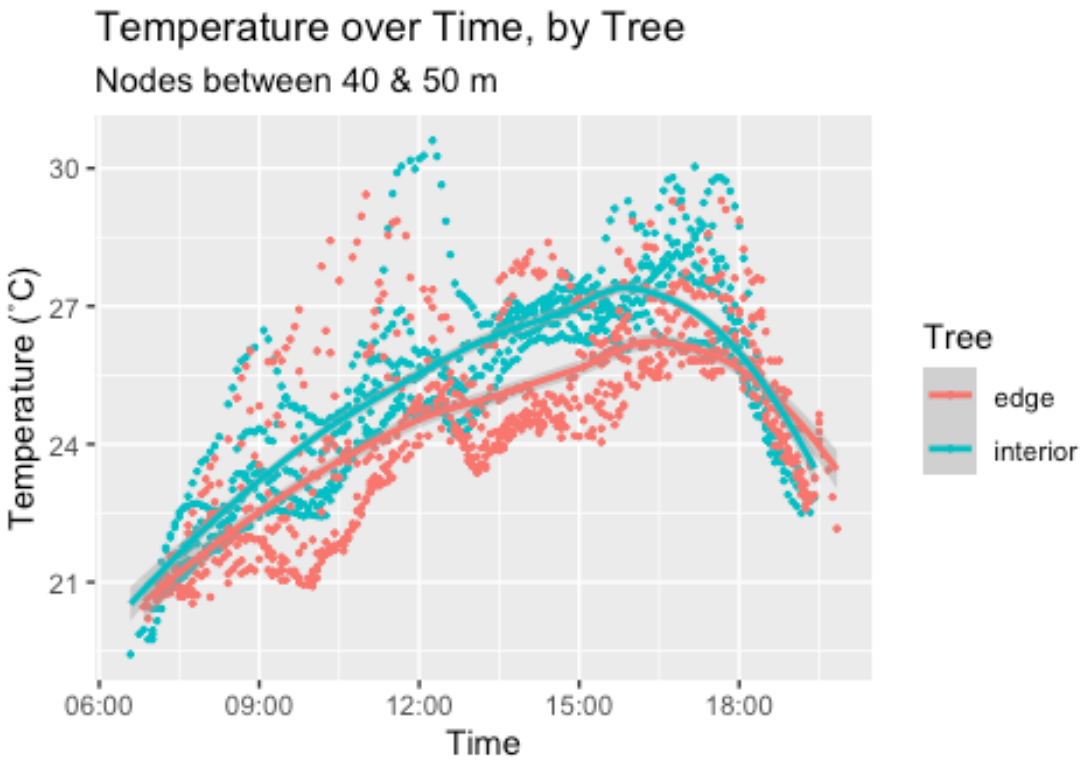
May 22, in contrast, was a cold day. Here, the trends are reversed, with the highest node recording the lowest temperatures. On that day, the median humidity that day was 96.5%. This indicates that the treetops were probably in a layer of fog which decreased the temperature.



4.3 Finding 3

This finding supports evidence for the existence of a second tree. The paper only ever mentions one tree, but the dataset categorizes measurements as belonging to either the "edge" or "interior" tree. It isn't clear if these refer to physical trees or the routing tree. To test this, I looked for differences in recordings of sensors in each tree. In particular, I tracked temperatures from 10 sensors at similar heights, sampling 800 points from each tree. (This data all came from the logger.)

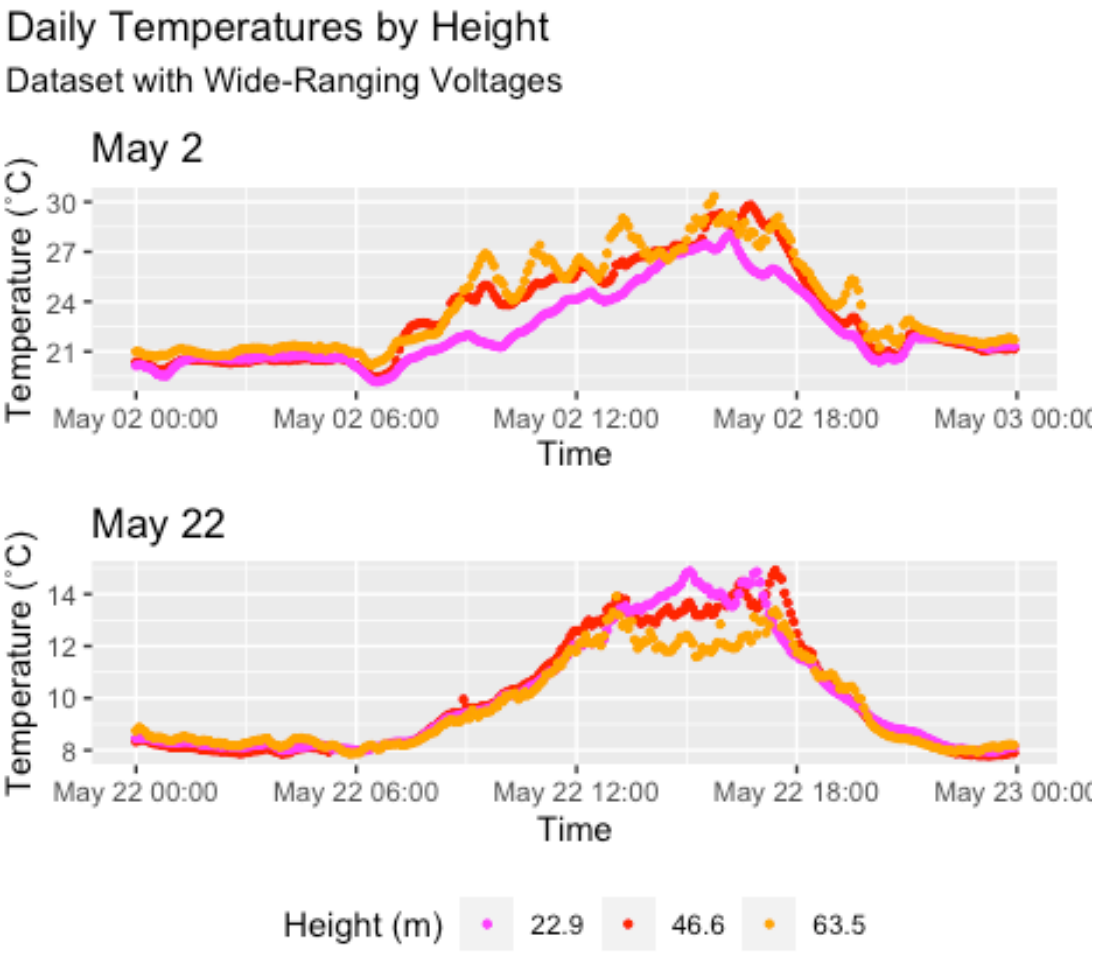
In the plot below, points from the interior tree tend to have higher temperatures than those from the edge tree. I used local polynomial regression smoothing to fit the data to curves. For most times in the day, these curves were far outside of each other's confidence intervals. Because the sensors have essentially the same height, we would expect their predicted temperatures should be equal if they came from the same physical tree. This leads me to believe that they came from two separate trees.



4.4 Stability Check

For the stability check, I reproduce Finding 2 with the data set aside in Section 2.2.4. Recall that this data has voltages between 2.1 and 3.3 V, rather than between 2.4 and 3.0 V. (I did all subsequent data cleaning on this dataset as well.) I chose Finding 2 because Figure 6 in the paper warned that faulty voltages produced temperature outliers.

In spite of this warning, the plots look all but identical to those in Finding 2. This may be due to our extensive data cleaning on the temperature data. Indeed, the dataset used here contains very few voltages that are outside the [2.4, 3.0] range.



5. Discussion

The dataset size was not terribly restrictive, but it did cause some trouble. For example, I wanted to remove duplicate readings by looping through the logger dataset and removing every row from the network dataset that shared the same (epoch, node ID) pair. I couldn't get this to work using apply(), so I tried using a for loop instead. This solution technically worked, but was not feasibly computable on our large dataset.

While we are able to draw interesting conclusions from this data, it is not a flawless representation of reality. The faithfulness of this representation is dependent on our recording mechanisms, which were far from perfect. After calibrating the devices beforehand, the researchers still observed a mean deviation in temperature of 0.18°C and 0.6% relative humidity. Moreover, the data was so messy that I imagine there were noisy outliers I neglected to remove. And even if the sensors were spot-on, our data would only provide a partial picture of the microclimate.

The next of the three realms for us to consider is modeling. In Finding 3 I used local polynomial regression smoothing to get a curve approximation for the relationship

between temperature and time. These results led me to conclude there were was strong evidence of a physical tree. The model uses a representation of reality, data, to enable us to better understand the true reality itself – whether our data represents one or two trees.

Data analysis and modeling should impact decisions on future data. Here, we hope to generalize our findings on the microclimate around this tree to that around other redwood trees. The authors mention in Section 6.2 that these temporal and spatial gradients can be used to validate biological theories. For example, ecologists could use it to model the effect that microclimatic gradients have on the rate at which sap flows through of a tree.

6. Conclusion

In this report, I explored data from a wireless sensor network that monitored the environmental dynamics around a redwood tree. Around the tree, a drop of 1˚C is typically causes a 5% rise in humidity. On chillier days, the tree gets progressively colder towards the top; the opposite is true on hotter days, where higher positions are as much as 5˚C warmer. These positions always receive more light. I also showed that the network and logger datasets contain duplicates, and the logger likely records two separate redwoods.

Our overarching goal in exploring this dataset was to better understand the dynamics of the microclimate surrounding a redwood tree. Tolle et al.'s network of sensors enabled us to observe environmental data over a 7 week span and 70 meter tree. I took advantage of this unique setup by analyzing data over temporal and spatial gradients. Overall, my insights elucidated ways in which environmental factors like temperature, humidity, and light vary around a redwood tree.

7. Academic Honesty Statement

Bin:

I pledge to uphold the highest standards of academic integrity throughout Stat 215 and beyond. Any contributions from classmates and other sources will be properly citied. It is important in any environment to give credit for people's work, but especially so in an academic setting, where one's contributions determine their success. Academic integrity is also beneficial to oneself. If someone is struggling on a problem, it is unethical and intellectually lazy to steal another person's efforts instead of working it through.

8. Bibliography

- Historical Temperature Data - UCCE Sonoma County. http://cesonoma.ucanr.edu/viticulture717/Viticulture_Newsletter/January_2011/Historical_Temperature_Data/. Accessed 14 Sept. 2021.
- Tolle, Gilman, et al. "A Macroscope in the Redwoods." Proceedings of the 3rd International Conference on Embedded Networked Sensor Systems - SenSys '05, ACM Press, 2005, p. 51. DOI.org (Crossref), https://doi.org/10.1145/1098918.1098925.