Project 1 Redwood Data Report

William Tirone (ID: 2774025) and Natalie Smith (ID: 2819547)

March 2, 2024

1 Summary of Tolle et. al.

With the desire to implement new technology with the capacity for dense temporal and spatial monitoring, Tolle et. al chose to investigate the day in the life of a single redwood tree across a month-long period. After the data was collected, the researchers used multidimensional analyses to reveal trends and gradients in the data across the variables of time, height, and value of readings (relative humidity, temperature, and photosynthetically active solar radiation (PAR)). Part of the main conclusion states the degree of success of the monitoring systems: the motes are effective in performing dense temporal and spatial monitoring in a biological setting. With these motes, this study was further able to show precisely the variation in temperature, humidity, and sunlight throughout different parts of the tree and at different times of the day, on a day-to-day basis. The researchers were able to conclude that the bottom of the tree is colder and more humid, on average, when compared to the top of the tree.

While all sensors were calibrated before the study began, the researchers did not take into account the effect of the extra testing for calibration purposes on battery life. Though some sensors were never in the normal range upon being placed on the tree, several had bad readings before dying, including on 5/26 when the data logs filled up. Outliers were removed for nonsensical data points and for when the voltage was above/below a certain threshold – a factor that indicated imminent battery death. Nevertheless, a plethora of meaningful data was still received by this wireless day-in-the-life monitoring, with over 800,000 measurements recorded. From this study, the technology shows its usefulness and effectiveness, demonstrating promise for the future of further environmental studies.

1.1 Sensor Deployment Summary

Sensor deployment covers placement on the tree, hardware and network architecture, and methodology. Time, vertical distance, angular location, and radial distance combined led to the specific placement of an individual sensor on the tree. Sensors were sampled every 5 minutes over the course of 44 days in the early summer, as summer contains the "most dynamic microclimatic variation." They were placed 15m from ground level to 70m from ground level with about 2m spacing between each node; 15m was the minimum distance since most foliage started at that point. The "west side of the tree had a thicker canopy and provided the most buffering against environmental effects," thus the west side of the tree was used for placement to shield the sensors and the casing from rain and adverse weather. The range of placement was 0.1 to 1m from the trunk to solely measure environmental effects on the trunk itself.

1.2 Measurements

The measured parameters chosen were "driven by biological requirements" and included temperature, humidity, and light levels. Light levels were measured using direct and ambient levels of photosynthetically active radiation. Temperature (+/- 0.5%) and humidity (+/-3.5 %) were measured by Sensirion SHT11 sensors, and Hamamatsu S1087 photodiodes measured incident and reflected PAR. The package for deployment that contained the sensors consisted of a mote, battery, and two sensor boards fit into a sealed cylindrical enclosure made of high-density white plastic. Data from the network of nodes were collected

and stored in a local database, then "transmitted over a cellular modem to an offsite database." Calibration of the sensors was performed on a rooftop and then secondarily in a contained environmental chamber to allow the sensors to be exposed to a wide range of phenomena before being deployed. After calibration, the nodes were taken directly to the Grove of the Old Trees in Sonoma, CA, where they were installed and then verified to be receiving data. Then, when the measurement window of 44 days was over, the nodes were taken down and dismantled to download the data log. While some nodes were destroyed by weather, most were not. The data in *sonoma-data-log.csv* is the data obtained by this process after deployment, while *sonoma-data-net.csv* was "retrieved over the wireless network" during the 44 days of the study.

2 Data Cleaning

2.1 Data Scaling and Range

In the two provided data sets, *data-log* measures voltage in volts, whereas *data-net* appears to be 100x the amount of voltage. Since the sensors have a maximum range up to 10 volts, from considering the scale sheets,[1][2] and consulting the orders of magnitude for voltage,[3] we believe the measurement for voltage *data-net* is in centivolts. Thus, we divided the latter data by 100 to maintain the same range. Additionally, to remain consistent with *Tolle*, we divided the depths of the nodes by 100 to convert them to meters. Figure 1 illustrates the need to change the range. To further immerse the reader, we color-matched Figure 1 to an image from the Grove of Old Trees.[4] Notably, one of the voltages from the incorrect range is very far from the majority of the data. Even after adjusting the scale, this needed to be removed as it far exceeds the specifications for the sensor.

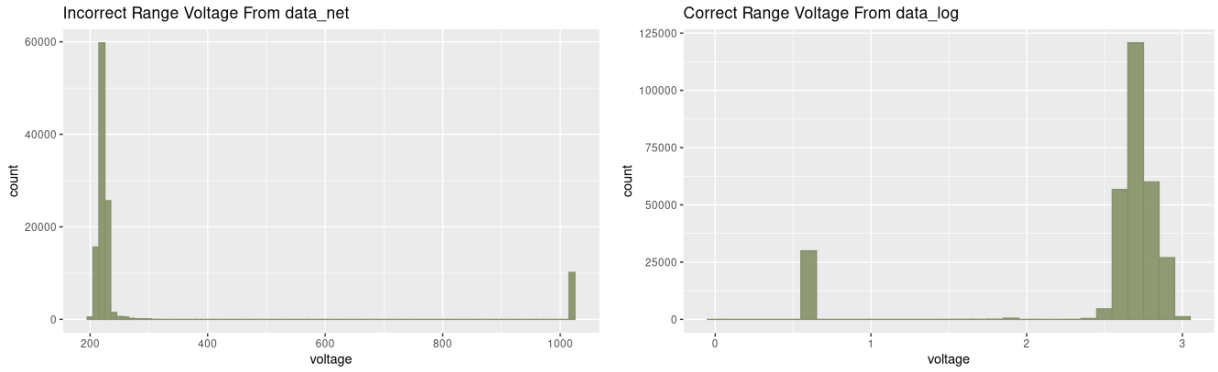

Figure 1: Incorrect and Correct Voltages

2.2 Removing NA Values

To remove missing data, we simply removed rows that had NA values for humidity, temperature, and both PAR measurements. For each measurement that was removed, every measurement column had NA values, indicating the sensor broke or stopped working. There were 12,532 rows with NA values. Additionally, per our classmate's discussion post on Ed,[5] we saw that the dates for the data are incorrect. Further, the *sonoma-dates* file was not in a manageable format, at least with R. The data appeared to be stored in variables that, while possible to work with, would have required manipulation. To get around this, we generated the data in R using seq(), then joined it to our main data set. From here, we threw away the original "result_time" variable. Result_time + epoch, as our classmate noticed, was not even consistent across *data-log* and *data-net*, and the time was not measured in 5-minute intervals as the paper stated it should be. Finally, with our intermittent cleaning task out of the way, we could correctly note that the range of dates for the NA data was from April 30^{th}, 2004 at 8:05 A.M. to May 29^{th}, 2004, at 3:35 A.M.

2.3 Mote Location Data

For the mote-location data, we left-joined the data on the nodeid / ID column to add 4 new variables to our total data: Height, Direc, Dist, and Tree. However, upon comparing the number of unique IDs in the location data to the total data, we noted we had 73 unique IDs in the total data but 80 in the location data. This difference could represent the sensors that broke or did not transmit data during the study.

2.4 Outlier Removal

We start our outlier removal by only keeping observations with voltage ≥ 2.4 and ≤ 5.5. The authors noted that a voltage level below 2.4 indicated a dying battery, and high voltages led to inaccurate readings. We also observed unusual data with these measurements and agree with the authors' decision to remove them. By removing these voltages before cleaning the rest of the data, we believe we will not have to manually remove as many values. Next, we removed values with humidity ≤ 0 and ≥ 100, as these are not possible.

To cross-validate temperature, we consulted historical weather data[6] from Santa Rosa, the nearest major city to the Grove of Old Trees to see if the temperatures measured in the data were at least plausible. The range from historical data was 0° F to 94° F, or 0°C to 34.4°C. The city of Santa Rosa may have very different temperature ranges from the forest, but it is the closest historical data we can find to give us a rough sense of what may be appropriate data to keep. We decided to remove any measurements ≥ 35°C by considering Figure 2, which removed a total of 20 rows of data.

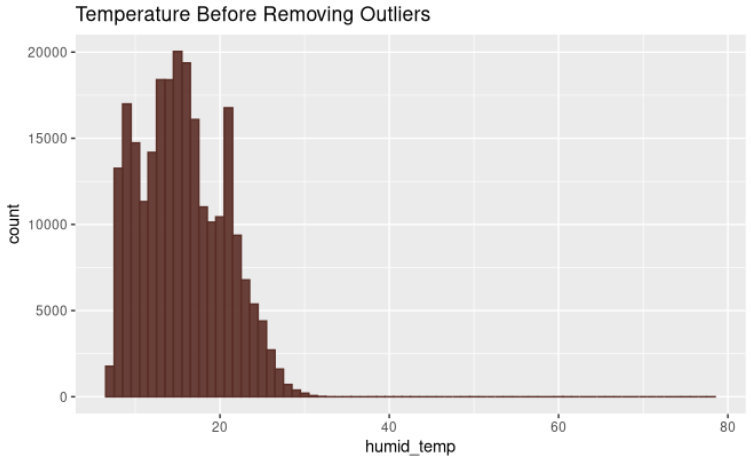

Figure 2: Temperature Readings

For both Incident and Reflected PAR, we divided by 54 to end up with the correct units, $\mu mol/m^2/s$, which agrees with the scale of the x-axes plots on p. 7 of Tolle et. al. This indicates the measurements were in Lux and had to be converted. A significant portion of the readings for Incident PAR is 0, which makes sense at nighttime. However, 16,529 of the measurements with 0 incident par are from 7 A.M. to 7 P.M. (Figure 3), which is most likely inaccurate since we would expect even a tiny amount of sunlight to make it to the sensors. Thus, we removed these. However, for the reflected PAR measurement, the same consideration above yielded about 65,000 rows, which seems to be too much data to remove entirely. As the bottom sensor will most likely receive less light, it is possible these values are appropriate. To handle this uncertainty, we left them in but will note when they are removed for plotting purposes.

2.5 Miscellaneous Data Cleaning and Removal

We noticed other unusual or incorrect values, similar to the bad date range from the NA values. For example, there are three values of "parent" that are 65535, the largest 16-bit unsigned integer. Though these have "normal" nodeids and measurements, we decided to remove them since the data appear untrustworthy,

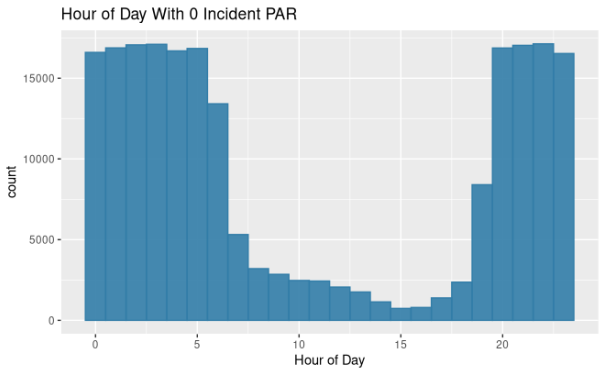

Figure 3: Considering PAR measurements and hour of day

but perhaps this was just an error in the nodeid assignment stage by the research team. We also noticed that some of the observations with NA values had a "depth" reading of 255, which is the largest 8-bit integer. This is not a possible measurement, since the depth is between 0.1m and 1m, but also may say something about the cause of the NA values. It seems possible that there were issues resulting from the database creation and setup.

3 Data Exploration

3.1 Scatterplots

Upon creating scatter lots for a variety of values, we found a few interesting relationships. Both plots consider data for the whole time period of the study. We created a categorical variable for day or night, with day being between 7 A.M. and 7 P.M., then used this to color the points for both of our plots (see Fig. 4). The majority of the data for temperature vs. humidity is clumped together, though in question 4, we attempt to cluster this data. For the most part, temperatures and humidity tend to be higher during the day, though there is still considerable overlap. There are also a handful of outliers with high humidity and temperature that occur at night, which could be useful to investigate further for either data removal purposes or for examining special weather patterns.

We also looked into the relationship between voltage and Incident PAR (Fig. 5), since Tolle et. al. discussed the importance of voltages in producing reliable measurements. There is a very interesting "hole" of data missing from the hamatop measurements during the day, from 2.4 to 2.6 volts, and a smaller piece missing where voltages are about ≤ 3.2 (recall that we removed measurements with voltages ≥ 5.5 or ≤ 2.4). This could be the result of our previous cleaning efforts, but indicates that the truly safe range of voltages in terms of having reliable data may be 2.6 to 2.9. Further, there are several nighttime points with hamatop readings from 2000 to 3000, with voltages from 2.8 to 3. These could be bad data points, but might also be some kind of interesting solar event that resulted in unexpectedly high nighttime measurements.

3.2 Correlation

There are three predictors that appear to have a relationship with incident PAR. One of the predictors is reflected PAR, which makes sense, as both are used to measure sunlight (moderate positive correlation; r = 0.5). Incident PAR also shows some association with height (moderately weak positive relationship; r = 0.2). There is also a weak positive relationship between incident PAR and temperature (r = 0.2).

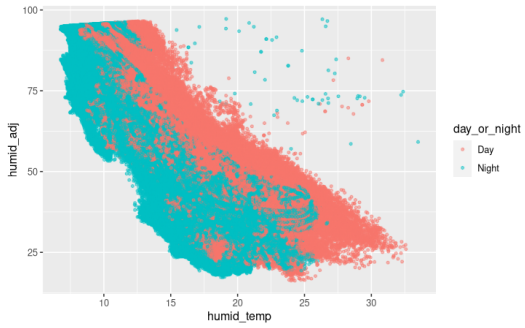

Figure 4: Temperature and Humidity Scatterplot, Colored by Time of Day

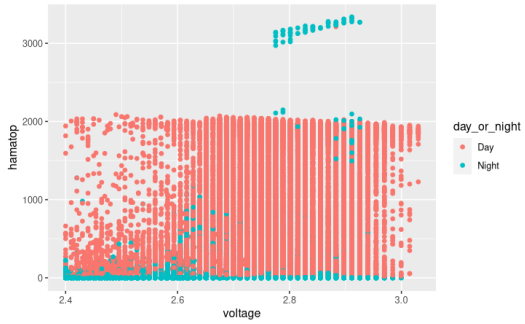

Figure 5: Incident PAR Plotted Against Voltage, Colored by Time of Day

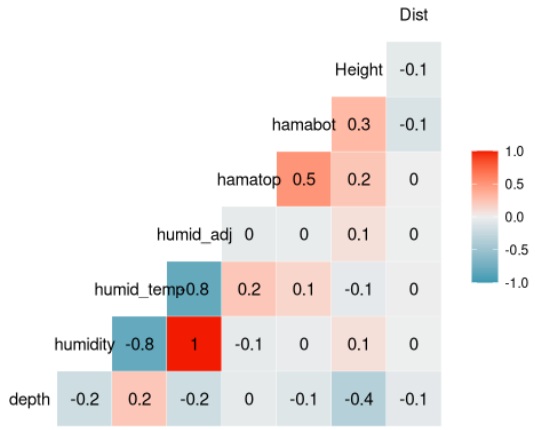

Figure 6: Variable Correlations

3.3 Dimensional Plotting

First, to plot with height as the color mapping, we divided height into 3 bins to roughly consider the bottom, middle, and top of the tree. We wanted to investigate the data using the date as well as the hour of the day for our independent variable, for comparison purposes (Fig. 7). Since the data has been heavily cleaned at this point, extreme temperature values were removed, so we observe temperatures from 0 to about 33. Humidity can take all values from 0 to 100. Hamatop (Incident PAR) and hamabot (Reflected PAR) take values from 0 to 3500 and 0 to 160 respectively. In all the plots, on average, the bottom of the tree sees less extreme values: lower temperatures, lower hamabot, and lower hamatop. Humidity, though, at the bottom of the tree was in line with the middle and top of the tree. Tolle et. al. noted that this was because the tree shielded the bottom from extreme weather effects. Temperature and humidity tend to be higher at the top of the tree from early May to early June.

Reflected PAR has an interesting presence of higher-valued readings from the bottom and middle of the tree until about May 10th, where the higher-valued readings are almost exclusively from the top of the tree (Fig. 7, bottom left). This phenomenon was also observed for reflected PAR (Fig. 7, bottom right).

Hourly, it is interesting to see that different nodes throughout the top of the tree had very different measurements despite being relatively close to each other (Fig. 8). Temperature varied by 3^oC, and humidity by about 13%. However, for these two measurements, the middle of the tree shows less variance, so if we wanted to draw inferences based on the data, it might be a good idea to use these measurements to remove some variation. For hamatop and hamabot, it looks like the top of the tree absorbed the large majority of the sunlight, with the middle and bottom receiving almost none (which is expected for a tall tree in a forest). There is still a large amount of variance here, though, among the different nodes.

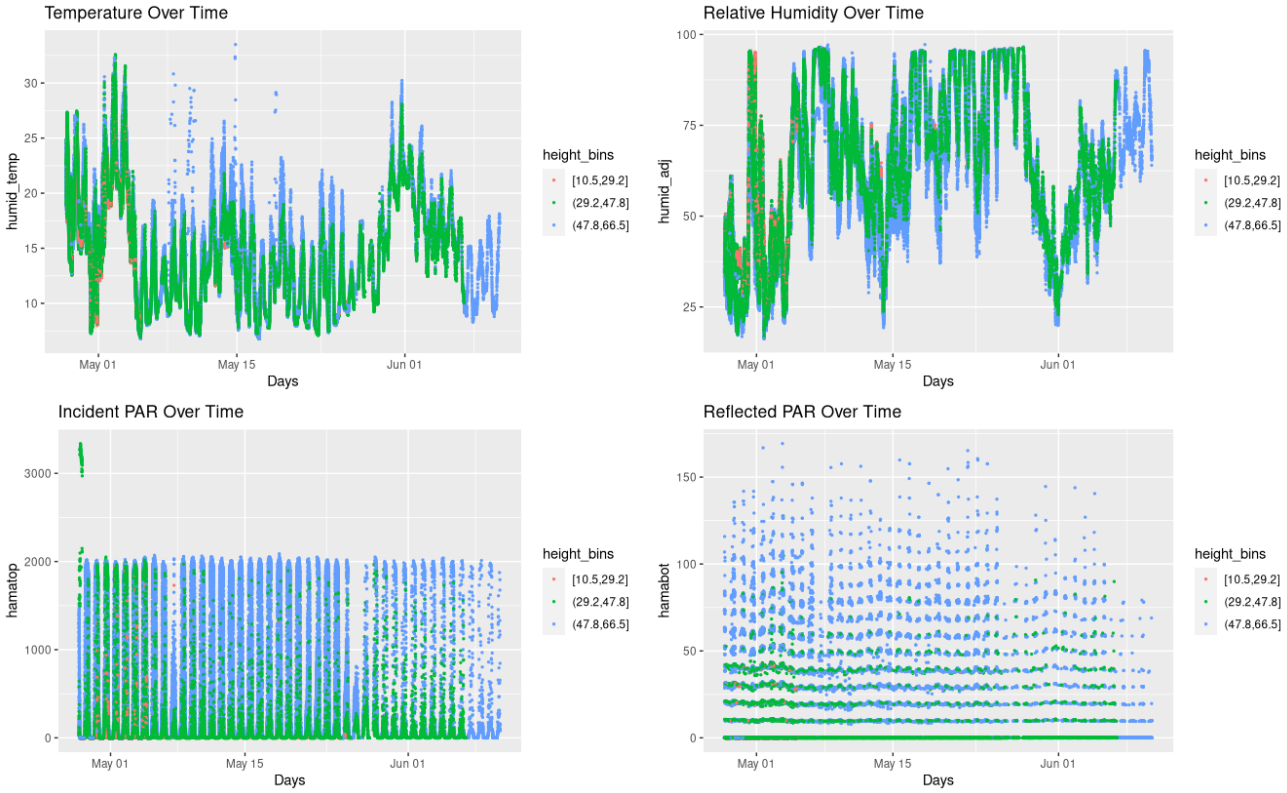

Figure 7: Temperature, Humidity, Incident PAR, and Reflected PAR by Days

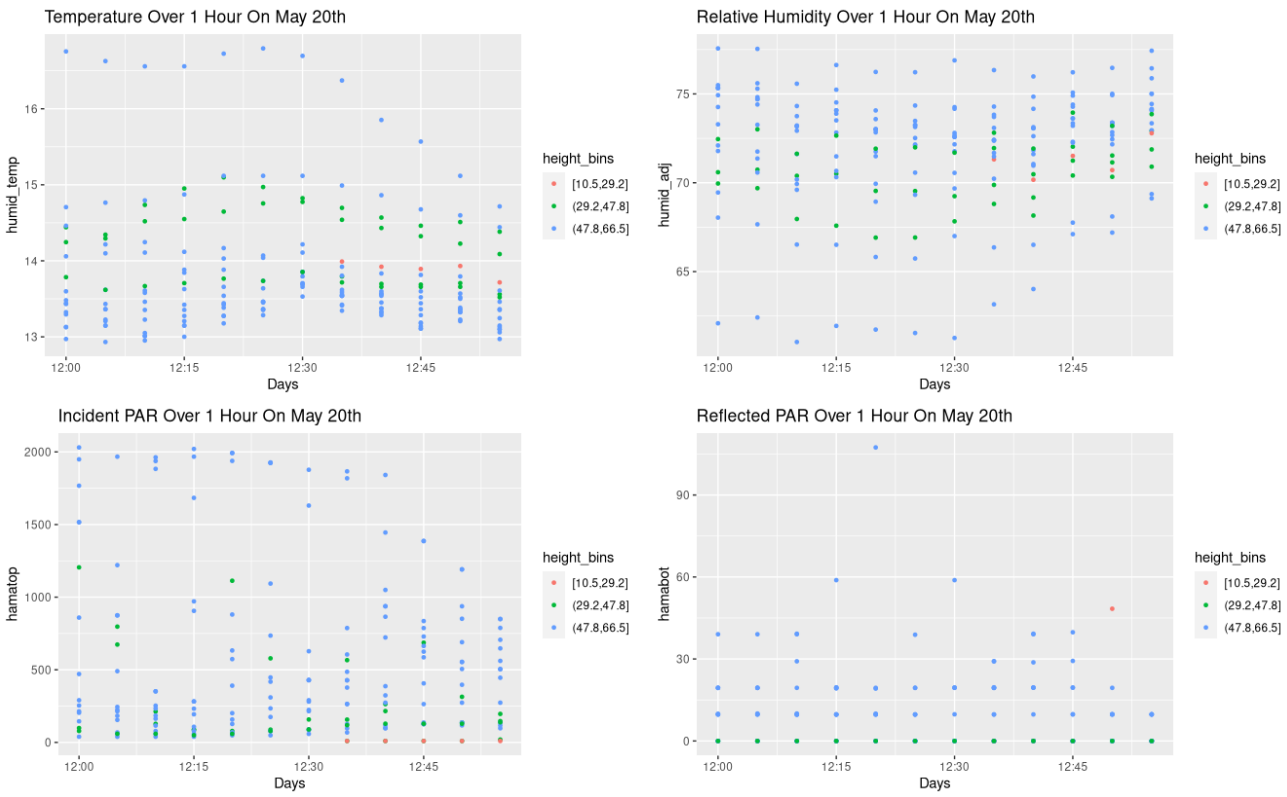

Figure 8: Temperature, Humidity, Incident PAR, and Reflected PAR Over 1 Hour

3.4 PCA

As we have a quite number of variables, dimention reduction is in our best interest to observe the underlying trends in the data. First, we created a screeplot to determine how many PCs would be useful to retain, and how much information is provided from retaining these PCs. From looking at the screeplot's elbows, we should keep around 4, 7, or 9 PCs. Another method to choose how many PCs to retain is more computational: look at which eigenvalue gets us to a value that is greater than the average of the eigenvalues, which we found to be the 4th eigenvalue. Ideally, for a low-dimensional representation, we would choose to keep 4 PCs. For the ideal 2-dimensional representation needed for ease of viewing, we plotted the first two PCs on the biplot in Figure 9. The biplot shows the PC directions for each of the variables inspected. As the first two PCs cover around 50 percent of the variation in the data, we can still gain valuable insight into the ways in which the data vary.

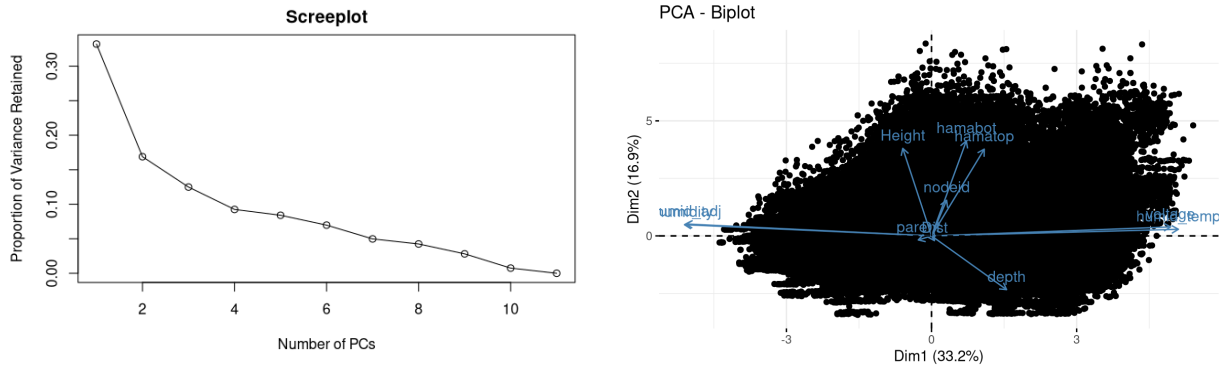

Figure 9: PCA

4 Interesting Findings

4.1 Finding 1: K Means on Humidity and Temperature

Intrigued by our findings in 3.1, we wanted to try pulling a subset of time for temperature and humidity to cluster. In figure 10, we chose May 1st to be consistent with Tolle et al.'s analysis of this date, plotted the measurements, colored them by the time of day, and then clustered. Since clustering assignments depend on the initial assignment, we set nstart=25 to reduce this impact. About 1/6 of the points in cluster 1 are below the average silhouette width, and thus near the decision boundary, while only about 1/10 of the points in cluster 2 are near the boundary. A very small number of points have a negative score, indicating they may have been assigned to the wrong cluster. Overall, though, the average score of 0.75 indicates that the clustering works moderately well.

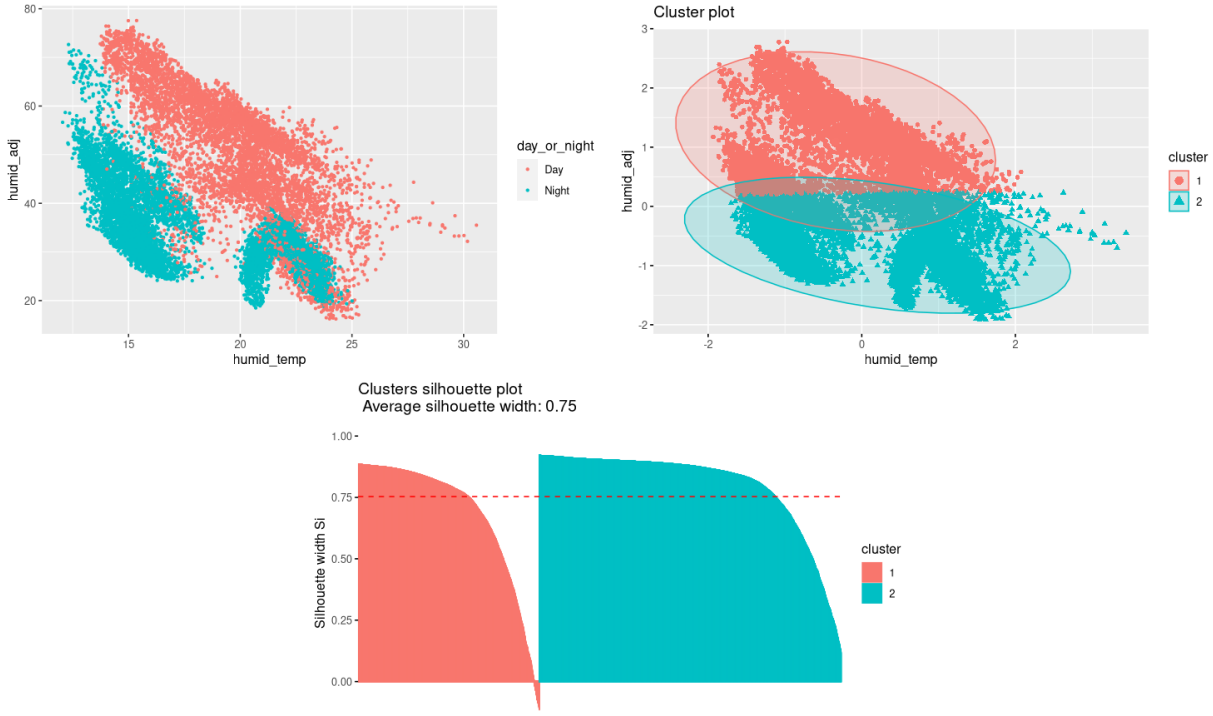

Figure 10: KMeans Clustering

4.2 Finding 2: How Tree Height Plays A Role in Value Measurements

Considering the negative relationship between temperature and humidity we saw in Figure 4, we were curious about how each tree height plays a role in the values. For temperature, humidity, incident PAR, and reflected PAR, we looked at the values outputted from the highest node on the tree and compared them against the lowest node on the tree by taking their difference (highest - lowest). We subsetted the data from May 31 to June 3rd to hone in on daily trends, and plotted this in figure 11.

For temperature, the tallest node can be upward of 2-4 degrees Celsius warmer than the node closest to the ground. For humidity, the lowest node can be 5 to 15 percentage points more humid than the tallest part of the tree. We continue to examine these relationships in section 5.3 of this report (Figure 14). In section 5.3, we only examine May 1st; however, we can clearly see how the top of the tree is warmer and less humid than the bottom of the tree.

We further examined the difference in sunlight between the top and bottom nodes of the tree. We can visually see how the top of the tree receives both more incident and reflected sunlight. What is interesting

to note here, however, is that just at around sunrise (approximately 6:30 am), the bottom tree node will have higher recorded values for Incident/Reflected PAR than the top node. We attribute this to the angle of the sun at sunrise hitting the bottom nodes first, but that is merely a conjecture, particularly as we hardly see the same trend at sunset, and minimally at best when looking at Reflected PAR alone.

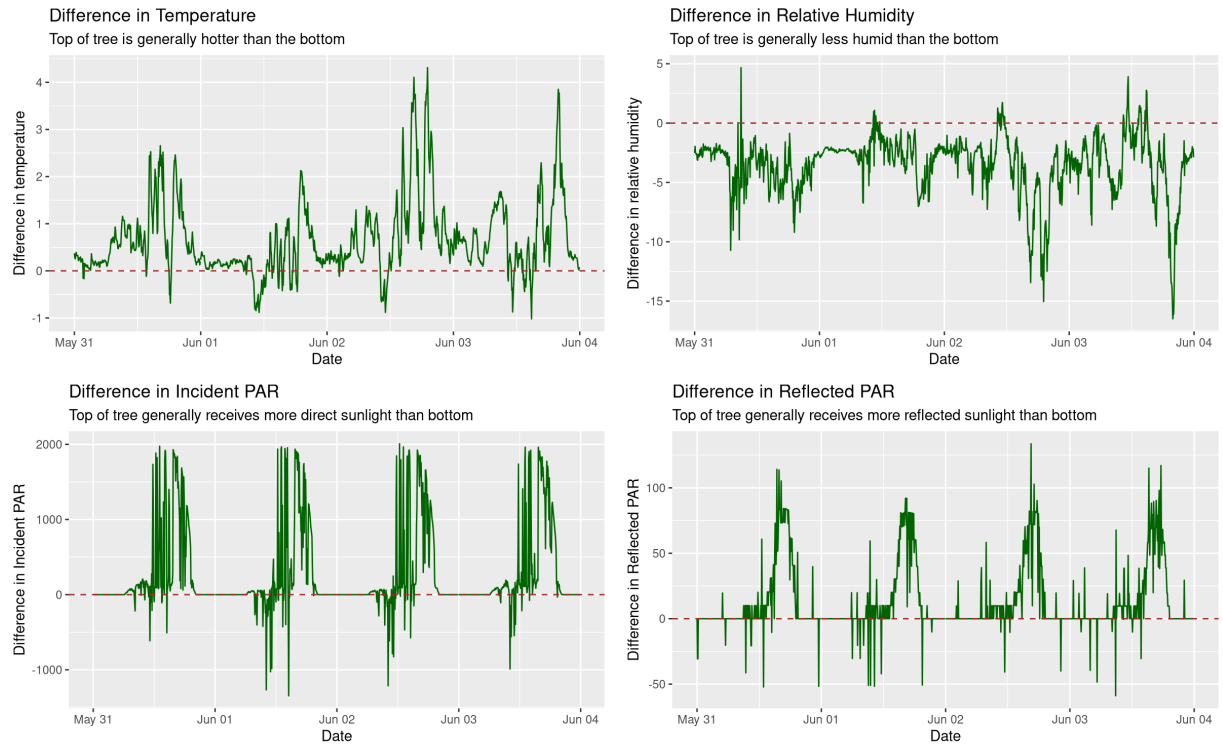

Figure 11: Exploring Values Between Top and Bottom of Tree

5 Graph Critique

5.1 Figure 3[a] Plot Critique

Figure 12 below shows the distributions of value readings log-transformed. The log-transformed data allows for a visual representation that is more symmetric, allowing us to better see the underlying distribution of our values of interest.

5.2 Figure 3[c], 3[d] Plot Improvement

The boxplots in 3[c] suggest that neither temperature nor relative humidity show trends based on node height, which as we've previously investigated, is seemingly false. When looking at incident PAR and reflected PAR, we see higher levels at higher node heights; however, the boxplot still shows that the majority of data for Incident and Reflected PAR are at low values, regardless of node height. The boxplots in 3[d] show the distributions of sensor reading differences from the mean (0) by node height. The same issue that arose in 3[c] is present here; to convey the trend by node height, we should present our findings here in a different manner.

As the research focuses on the "life of a tree," the plots should focus on how temperature, relative humidity, incident PAR, and Reflected PAR change by node height. These values are likely to be different, but minimally. So, using boxplots (which are used best to detect when differences are large) is not the best method to convey this idea; we believe Figure 14 below does a better job of characterizing this. However,

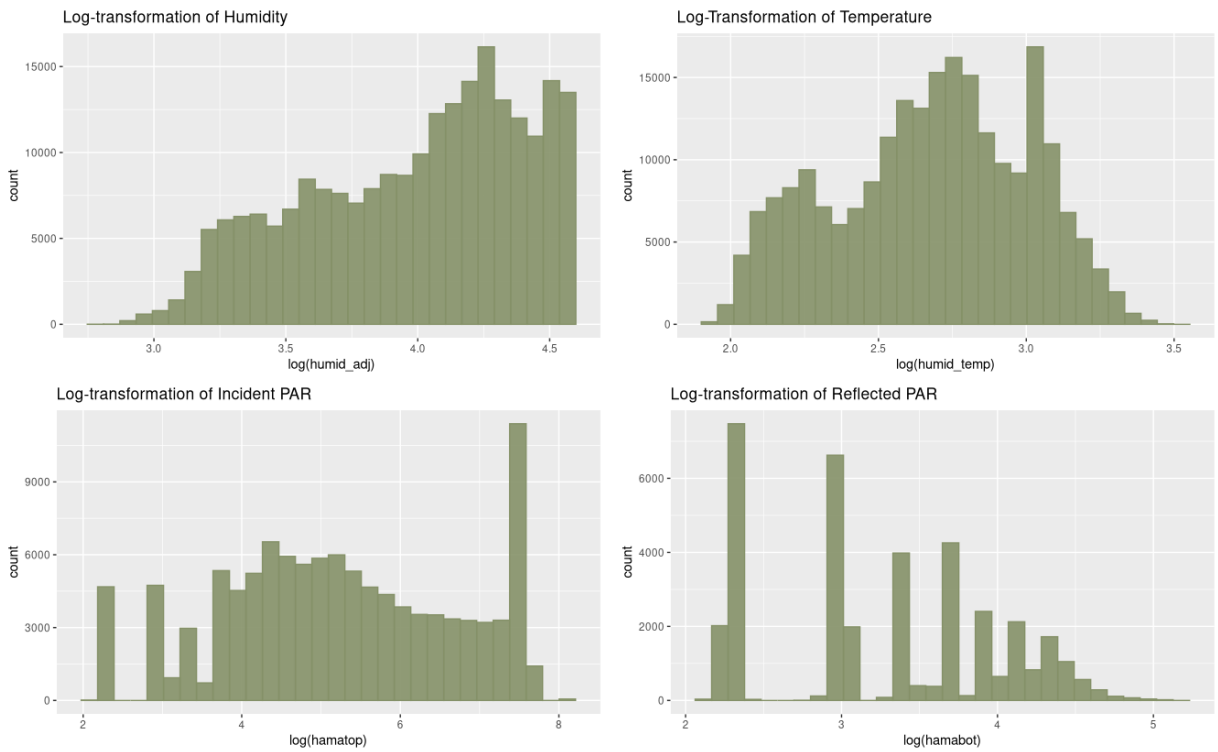

Figure 12: Log Transforms of Humidity, Temperature, Incident PAR, and Reflected PAR

to display the same data, we decided to bin our nodes by height into five sections and create boxplots regarding the distribution of that data. This brings a clearer picture of the range of values the notes receive without overwhelming the viewer.

The reason why boxplots aren't the best means to visualize this is throughout the 44-day period, the entire tree experiences humidity and temperature changes - the top of the tree on a colder day can produce the same reading as the bottom of the tree on a warmer day. Without grouping by any other variable (ex/ time of day, or a single day in the life), the range will likely remain the same, thereby not producing anything worthwhile.

Figure 13 shows our more informative boxplots, which retain the same amount of data (i.e., demonstrating the underlying distribution of values), but present it in a less overwhelming fashion. The boxplots for incident PAR and reflected PAR are not interesting as most of the data values are at 0 (no sunlight/night). We removed these values when recreating our plots to show the distribution of the data when there is sunlight.

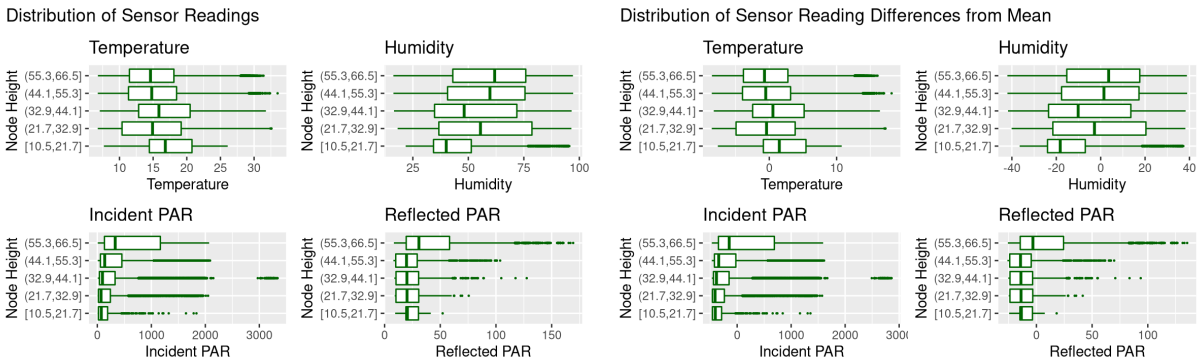

Figure 13: Plot Improvement

5.3 Figure 4 Plot Improvement

Generally, no line plot should have more than five colors; beyond that, the colors become too difficult to differentiate from each other, consequently serving no purpose. The two plots in the Tolle et al. paper have far too many lines that are indistinguishable from one another. The reader has to assume what each line serves. To improve this visualization, we added plot titles and a legend so the reader can distinguish the meaning behind the visualizations. In terms of the graphics, we only colored in the top node and the bottom node and made the remaining nodes a grey color, still showing the range in values produced. These changes are made in Figure 14.

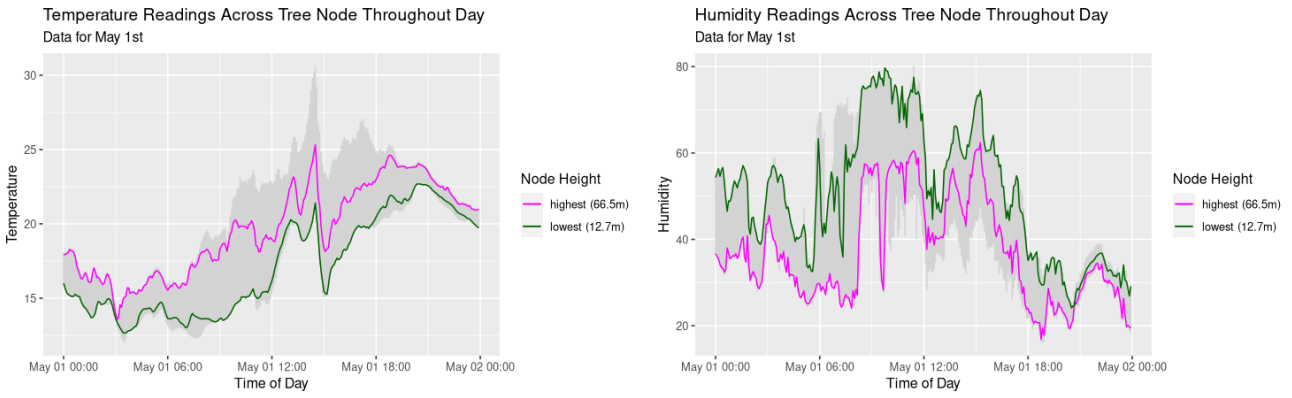

Figure 14: Improved Plots for Temp and Humidity Readings Colored By Node Height

The instantaneous reading plots (to the right of the temperature and humidity time plots in the paper, respectively) were confusing, mainly because one could not infer what the plot depicted without needing to inspect the paper. The triangles and corresponding colors don't signify much without a legend. Nonetheless, the plot does a good job of characterizing how the values change across node height at a single point in time. So, to improve these plots, we kept the underlying nature of the plot the same but included informative titles and a legend. We also scaled our axes so that the data fills up the span of the coordinate space. These changes are seen in Figure 15.

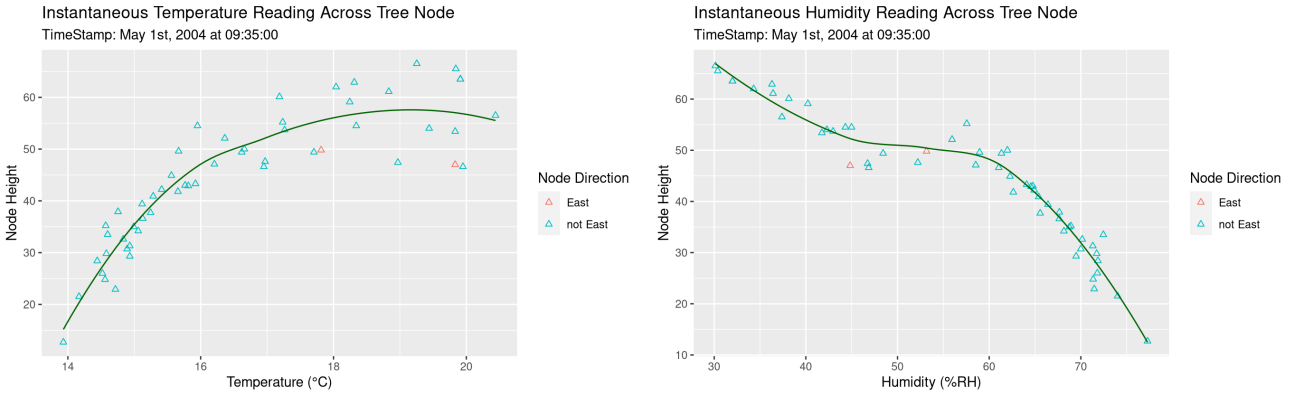

Figure 15: Improved Plots for Instantaneous Temp and Humidity Readings

5.4 Visualization of Log and Network Data

The visualizations in Figure 7 of Tolle et. al. provide information independently, but not through comparison (which is what Figure 7 aims to do). The last plot is mainly confusing, as the only x-axis measurement is day; however, the purpose of the bars seems to be to represent yield.

It is unclear how Tolle et. al. are reporting measurements vs. no measurements since there are 12,532 null measurements and 403,504 with readings. They note that the most common yield percentage is 0, though they don't specify what they consider "no reading." This could be data they rejected, which was then considered "no reading." but it is difficult to say. Conservatively, we decided to check the source of the null data rather than try to draw conclusions about the author's original intent and the yield. Our plot below allows for a quick comparison without having to jump around between 8 graphs to try to understand the yield from log or net. This is plotted in Figure 16 below.

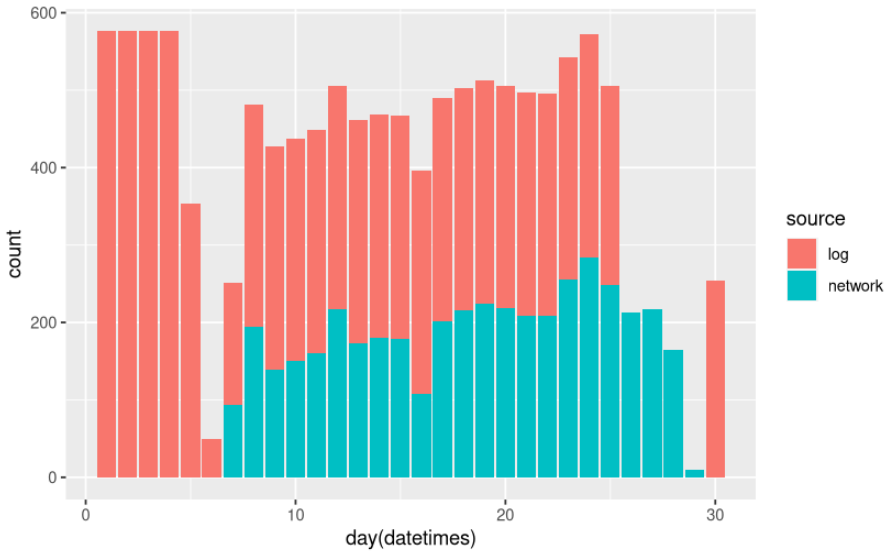

Figure 16: Missing Values From Log Data and Network Data

References

[1] https://sensirion.com/products/catalog/sht11/. Online; accessed 10/9/2022.

[2] https://docs.rs-online.com/6fa1/0900766b809880e5.pdf. Online; accessed 10/9/2022.

[3] https://en.wikipedia.org/wiki/voltage. Online; accessed 10/9/2022.

[4] Don Letherman. https://dons.zenfolio.com/p801593061/h25452cf9h3042827a. Online; accessed 10/9/2022.

[5] Fatima Alqabandi. Ed Discussion Post. https://edstem.org/us/courses/27257/discussion/1745672. Online; accessed 10/9/2022.

[6] https://www.wunderground.com/history/monthly/us/ca/santa-rosa/ksts/date/2004-6. Online; accessed 10/10/2022.