Project01

Jonathan Stuart (3032239913) and Nikhil Sakhamuri (3031843166)

3/25/2019

Data Collection

This was a case study of a wireless sensor network that recorded 44 days in the life of a 70-meter tall redwood tree from Tuesday, April 27th 2004 to Thursday, June 10th 2004. The readings were taken from a redwood tree located in the Grove of Old Trees in Sonoma, California. Sensors were taking readings every five minutes throughout these 44 days, leading to a total of 1.7 million data points over 33 motes. Each mote contained various sensors that measured four different data values throughout their deployment. The four values of interest were temperature, relative humidity, incident photosynthetically active radiation (PAR), and ambient PAR. After the conclusion of the study and analysis of the obtained data, several lessons were learned and several important takeaways were made. The first lesson that was learned was the extreme importance of being extremely precise when setting up delicate sensors. This is because even a slight variation in the position, angle, etc. of extremely sensitive sensors can cause unwanted variation in the data. This is what happened with the PAR sensors in this study; initially it was thought that fluctuations in the PAR readings were caused by differences in received sunlight caused by blocking foliage. However, it was later determined that the fluctuations were similar on similar days, meaning that the fluctuations were actually caused by slightly different orientations for each light sensor. A second lesson that was learned in the end was the importance of having a better apparatus for catching and handling failures. Of the 1.7 million data points that were supposed to be received, only 820,700 were usable for analysis. Loggers with larger memory space are one option to avoid this. In conclusion, the deployed macroscope of wireless sensors captured the complex environmental interactions of the microclimate surrounding a coastal redwood tree. This study affirmed the existence of spatial gradients in the microclimate around a redwood tree and captured enough data to track the changes in these gradients over time. This data can be extremely useful in helping to validate other biological theories. This can in turn lead to a better grasp of large-scale processes of carbon and water exchange within a forest ecosystem.

All the data was recorded during a 44 day period from Tuesday, April 27th 2004 to Thursday, June 10th, 2004. Measurements were taken every 5 minutes during this period. This came out to a hypothetical total of 50,540 real-world data points per mote, or a total of 1.7 million data points over 33 motes. However, only 820,700 points were usable. . The four variables of interest were temperature, relative humidity, incident photosynthetically active radiation (PAR), and ambient PAR. Temperature and relative humidity were measured because they are necessary variables when attempting to analyze transpiration patterns of redwood trees. Incident PAR provides information about the energy available to plants for photosynthesis and gives insight about the drivers for carbon balance in the forest. Additionally, satellite remote sensing measurements of the reflectance of the land surface can be validated by looking at the ratio of reflected to incident PAR. Initially, the scientists considered measuring total solar radiation and barometric pressure as well. However, both these measurements were decided against. Total solar radiation was left out because it required sensor was overly sensitive and PAR readings gave pretty much the same information. Barometric pressure was excluded because it was concluded there were no appreciable differences in barometric pressure across the height of the tree. The data was collected by deploying a macroscope of wireless sensors across a singular redwood tree. Starting at 15 meters from ground level and going up to 70 meters from ground level, nodes of the macroscope were placed on the tree with 2 meters of spacing between nodes. This spacing ensured that the gradient of each variable could be captured with adequate precision. Most of the sensors were placed on the west side of the tree because the west side had a thicker canopy which would help shield the nodes from direct environmental effects. Additionally, each node was placed between 0.1-1m from the trunk in order to ensure that the sensors would record only the microclimatic trends of the tree rather than the wider environment. In order to account for possible failures in the system to properly record and transmit the data readings, a local data logging system was used. The data logger recorded every reading taken by every query

before the readings were passed through the larger network. This local data logging system was used later to analyze the performance of the system as a whole. Sonoma-data-log.csv contains the data retrieved from the data logs after deployment while sonoma-data-net.csv contains only the data retrieved over the wireless network. Sonoma-data-net.csv has a much lower yield than Sonoma-data-log.csv; Sonoma-data-net.csv is missing an entire two-weeks of data, whereas that two-week period is present in Sonoma-data-log.csv.

Data Cleaning

Finding and Scaling Inconsistent Data
After viewing the histograms below (See Fig. 1), we notice that the voltage data is the only data that appears to be inconsistent between the net and log data sets. In general, the voltage values in the net data set appear to be about two orders of magnitude larger than the voltage values in the log data set. In order to convert the data to the same range, we used a vectorized operation to divide each of the net voltage values by 100, and then reassigned that scaled column to the original voltage column in the net data set.

Comments on Missing Data

In order to count the number of missing values in each of the datasets, we wrote a function that calls `is.na()` in a vectorized fashion over each of the columns in a given data set and sums the number of affirmative responses. Those sums are then stored in a vector and returned as an atomic vector of integer values. That is, we wrote a function to count the number of missing values in each column and store the sum in a new vector. We found that the net data set had the fewest number of missing values with 4263 observations returning a missing value in each of 5 columns. The log data set had roughly double that amount of missing values with 8270 observations returning a missing value in each of 5 columns. We counted missing values for the humidity, temperature, adjusted humidity and top and bottom PAR values (incident and reflected). When each variable returned a missing value, missing values were also returned for the 4 measured variables. All other variables in the data sets yielded no missing values.

With respect to the time period, we found that the log data initially gave no insight into when the missing values were recorded because the Result Time variable was a uniform value throughout the data set. Using Epoch values as a proxy for time values, we produced a visualization that shed some light on the missing values. We found that after the 2500th (approximately 05/08/2004) epoch up until the 8000th epoch approximately 05/25/2004), the missing values followed a relatively cyclic pattern before establishing a rhythm of constant increase. We also found that the number of net-collected missing values was roughly a third of the number of log-collected missing values during this same time period (See Figure 2).

Incorporating Location Data

After incorporating the location data into the net data file, there are now a total of 15 variables in our `new_net_data` dataframe.

Investigating and Removing Outliers

Examining the quantiles and studying the plots in Figures 3, 4 and 5, we came to the conclusion that we should eliminate humidity values over 100%, as was also mentioned in the Tolle, et al. paper. We also found that temperatures below zero and over 100 were likely outliers. We made these decisions with respect to the removal of outliers given out understanding of Sonoma County and the weather it experiences during the months when the experiment was conducted. We also applied the technique discussed within the Tolle, et al. paper whereby records were removed when that record's voltage fell outside of a predetermined range, even though this method is flawed, as discussed in our findings. This decision was made because it is reasonable to assume that sensors with peaking or failing voltage measurements were incapable of recording accurate data measurements, even if only some of the data measurements are compromised, as discussed in our findings.

Data Exploration

Pairwise Scatter Plots
Scatter plots of slected variables can be seen in Figures 6, 7and 8. In Figure 6 examining the relationship between temperature and depth of sensor placement, the plot indicates that as the depth of sensor placement increases, temperature drops. Figure 7 compares incident and reflected PAR in the context of depth of sensor placement. Here, we see that the smaller depth values showed greater correlations between reflected and incident PAR, indicating that sensors at a depth of 1 got the most reflected PAR as increases in incident PAR correlated positively with increases in reflected PAR. Senors at higher depths, however, showed less and less increase in relfected PAR per increase in incident PAR. Finally, in Figure 8 which examines the relationship between temperature and humidity, the plot clearly indicates that as humidity rises, temperature drops.

Fot the time period, we chose the time between epochs 3000 and 10000 as our exploratory data visualization indicated that epochs outside of that range showed uncharacteristic data patterns, especially in the range before epoch 3000.

Incident PAR
Hamabot and Hamatop are associated with relected and indident PAR, respectively. The paper disucsses how motes were able to measure photosynthetically active radiation both toward the top of the tree and toward the bottom of the tree. Motes placed further toward the top of the tree were able to measure incident PAR, and the radiation measured by these sensors came directly from the atmostpher. Sensors toward the bottom of the tree, however, mainly had access to reflected PAR. In this way, Hamatop is associated with Incident PAR and Hamabot is associated with Reflected PAR.

Temporal Trends

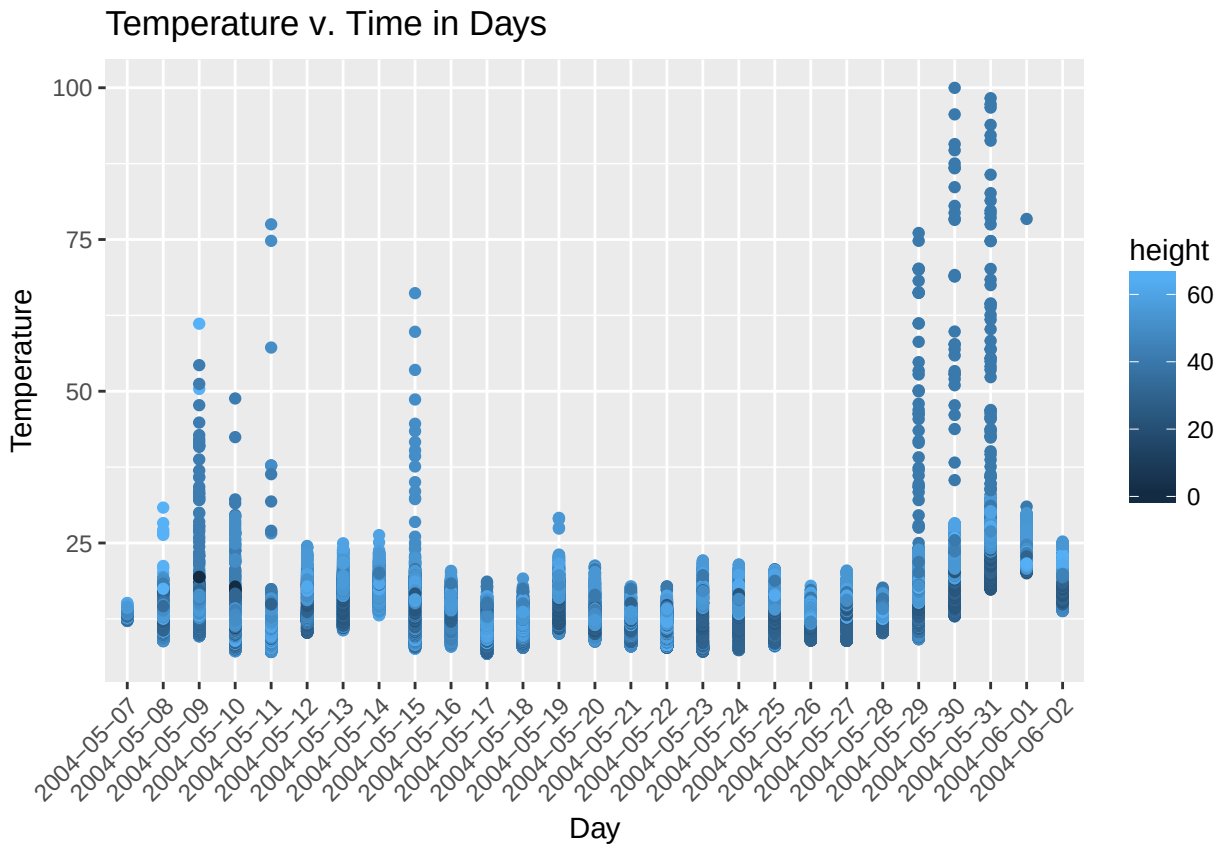
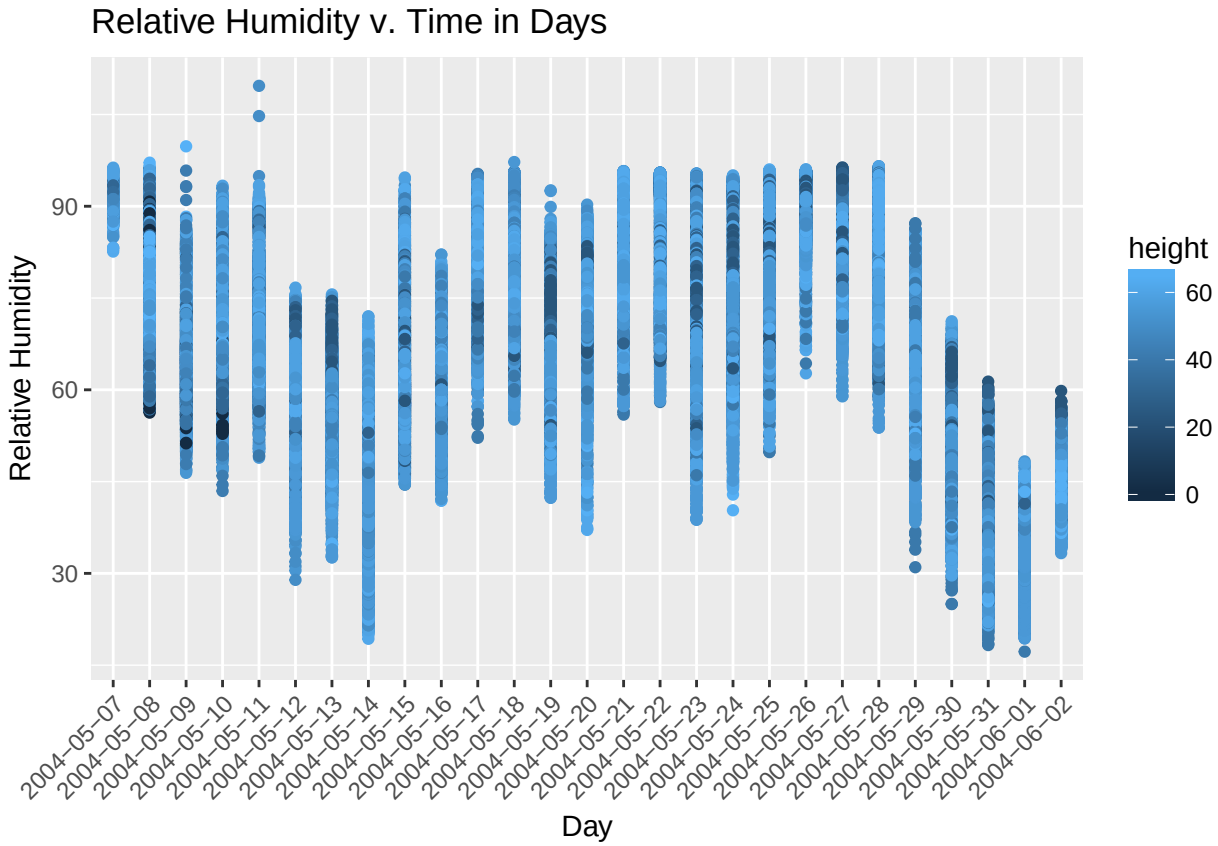
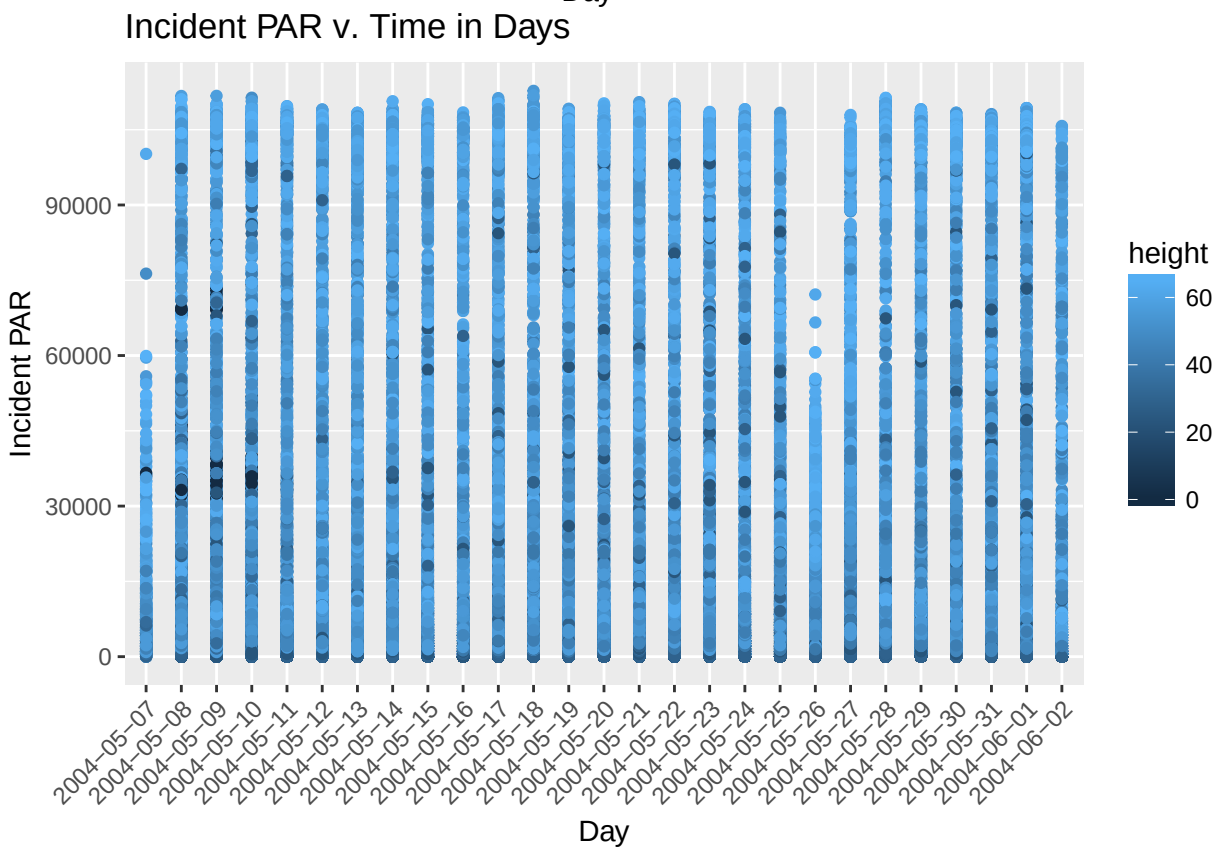


Figure 1: Temporal Trends - Temperature


Relative Humidity v. Time in Days
Relative Humidity
90
60
30
2004-05-07
2004-05-08
2004-05-09
2004-05-10
2004-05-11
2004-05-12
2004-05-13
2004-05-14
2004-05-15
2004-05-16
2004-05-17
2004-05-18
2004-05-19
2004-05-20
2004-05-21
2004-05-22
2004-05-23
2004-05-24
2004-05-25
2004-05-26
2004-05-27
2004-05-28
2004-05-29
2004-05-30
2004-05-31
2004-06-01
2004-06-02
Day
height
60
40
20
0


Incident PAR v. Time in Days
Incident PAR
90000
60000
30000
0
2004-05-07
2004-05-08
2004-05-09
2004-05-10
2004-05-11
2004-05-12
2004-05-13
2004-05-14
2004-05-15
2004-05-16
2004-05-17
2004-05-18
2004-05-19
2004-05-20
2004-05-21
2004-05-22
2004-05-23
2004-05-24
2004-05-25
2004-05-26
2004-05-27
2004-05-28
2004-05-29
2004-05-30
2004-05-31
2004-06-01
2004-06-02
Day
height
60
40
20
0

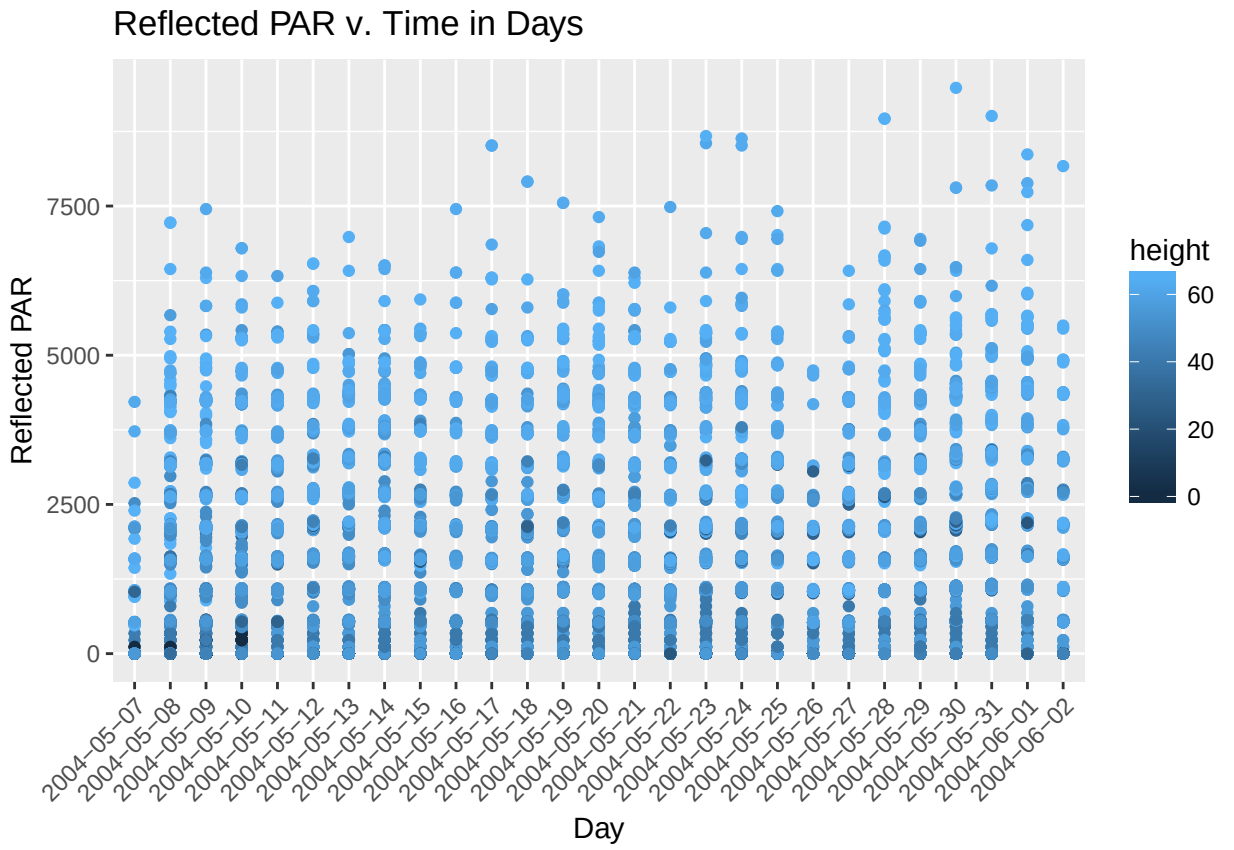


Figure 2: Temporal Trends - Reflected PAR

The plots of temporal trends expose interesting patterns and behavior within the data. The Reflected PAR plot, for example, clearly shows that for lower heights, darker blues, the data value was consistently lower, indicating that the lower sensors were exposed to less reflected sunlight. This makes reasonable sense, as we can expecte that the foliage of the tree would block sunlight and negatively impact the amount of sunlight that reached the sensors. With temperature and humidity data, these plots show an interesting cycle. Considering the humidity data, for example, we see a day to day change in the range of the temperature values. The temperature values, grouped tightly over a consistent range of about 50 degrees moved up and down, day to day. That is to say that the high and low of each day increased and decreased in tandem. From this, we can see that the least humid days occurred in mid-May and early June. Interestingly, carrying that analysis over to the temperature data, we can see that the hotter days occurred when humidity was lowest, an inverse relationship present in our analysis and mentioned elsewhere in this write up. Strangely, we see a large number of perturbed temperature measurements in early June, with some ranges on a given day spanning 75 degrees instead of the usual 15 or so degrees.

PC Analysis

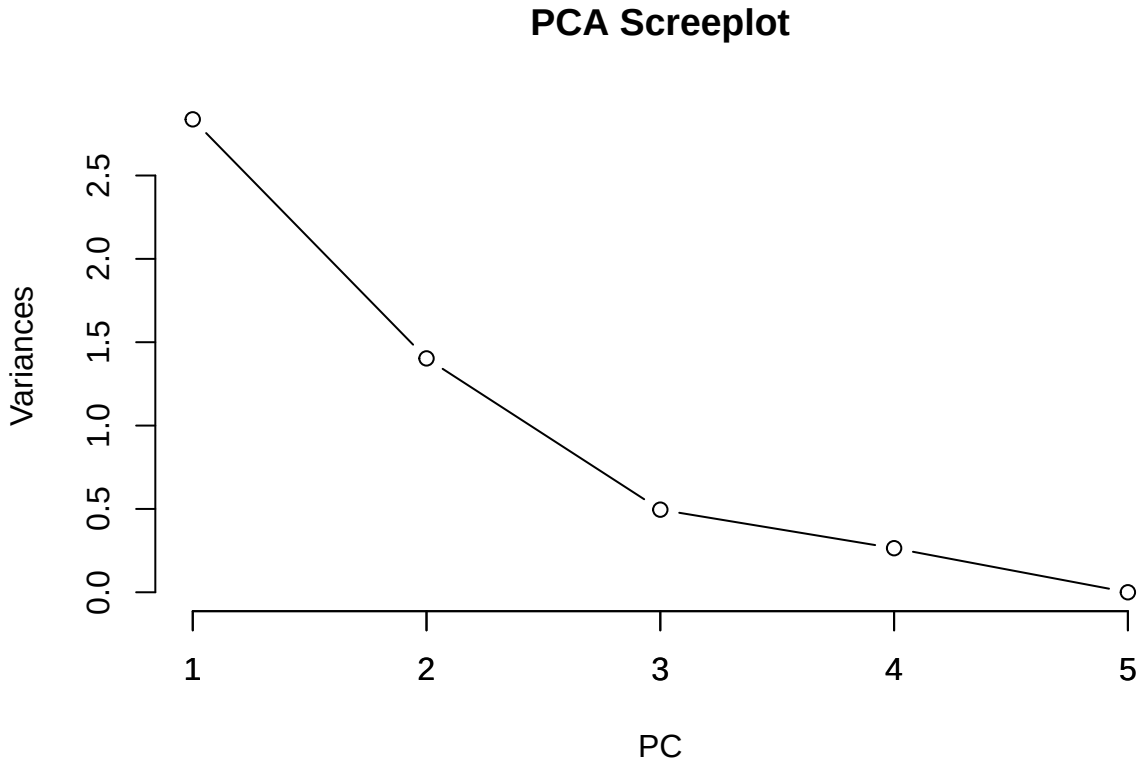


Figure 3: PCA Screeplot

Judging by the scree plot, we can conclude that this data can in fact be approximated by a lower dimensional representation. We use scree plots to help decide which principal components to keep for analysis, and we make that judgement based on the slop of the line connecting the principal components in the scree plot. the steeper the slope, the more telling the principal components. In general, the principal components connected by a line of similar, steeper slope are those we would wish to keep, discarding those that occur after the line starts to flatten. In our plot, the first "elbow" or "knee" occurs between the second and third principal components, indicating that the first and second principal components can be taken to represent the given data with reduced dimensionality.

Findings

Finding 1
One interesting finding that I found is with regards to the relationship between voltage present in the sensors and the humidity, incident PAR, and reflected PAR readings. In the Tolle, et al. paper, the relationship between voltage and temperature readings were discussed and it was found that the outliers in temperature readings were correlated with dips in the voltage readings. I was curious to see the relationship between voltage readings and the readings of humidity, incident PAR, and reflected PAR. First, I plotted four plots of all readings transmitted across the network with time on the x-axis and voltage, humidity, incident PAR, and reflected PAR on the y-axis respectively. I found no strong correlation between outliers in voltage and outliers in either of the other three readings. Wanting to go further, I then filtered all the network data to only include readings where the voltage of the sensor was between 2.0V and 3.0V. Interestingly, even after removing outlier voltage readings humidity, incident PAR, and reflected PAR maintained their full ranges of values respectively. This supported the conclusion after the first set of plots that there was no relationship between outlier voltage readings and outlier humidity, incident PAR, and reflected PAR readings.

This is important because in the outlier rejection portion of the report, the only real method of rejecting outliers was rejecting readings from sensors that had non-functional voltage readings. Therefore, only outlier temperature readings were rejected and outlier humidity, reflected PAR, and incident PAR readings were largely ignored. These findings illustrate the importance of finding a better method of finding outliers in the humidity, reflected PAR, and incident PAR readings. Once a better method is found, these outliers can also be rejected and the presented data can be more accurate with regards to its findings. The relevant graphs can be found in Figures 12-19.

Finding 2
Another interesting finding has to do with the microclimates that are discussed in the Tolle, et al. paper. Reading the paper and then analysing the data, I was struck by the idea that a single biological organism might both create and be impacted by multiple atmospheric environments. This is first and most clearly visible by the amount of photosynthetically active radiation available to the plant toward the top of its height and toward the bottom. This difference in available PAR is then mirrored in the temperature and humidity dynamics experienced by the tree depending on both the height and depth of consideration. For example, in the temporal trends plots, it is clear that points lower experience lower temperatures, higher humidities and lower incident PAR values, while it is equally clear that points higher experience an almost opposite dynamic. My major takeaway from the analysis of the data in this study is the dynamic and adaptive nature of large vegitation, including its ability for different parts of itself to be create and be exposed to different microclimates.

Graph Critique

Our log-transformed graphs of relected and incident PAR can be seen in Figures 20 and 21.

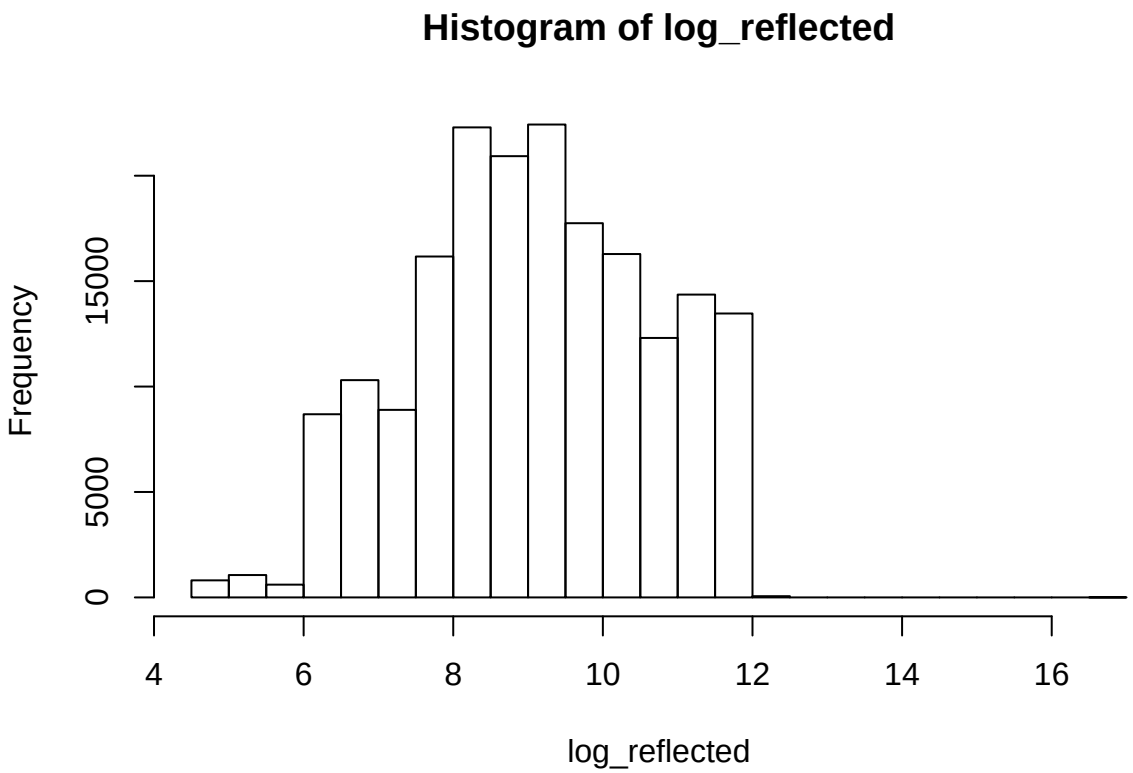


Figure 4: Log-transformed Reflected PAR

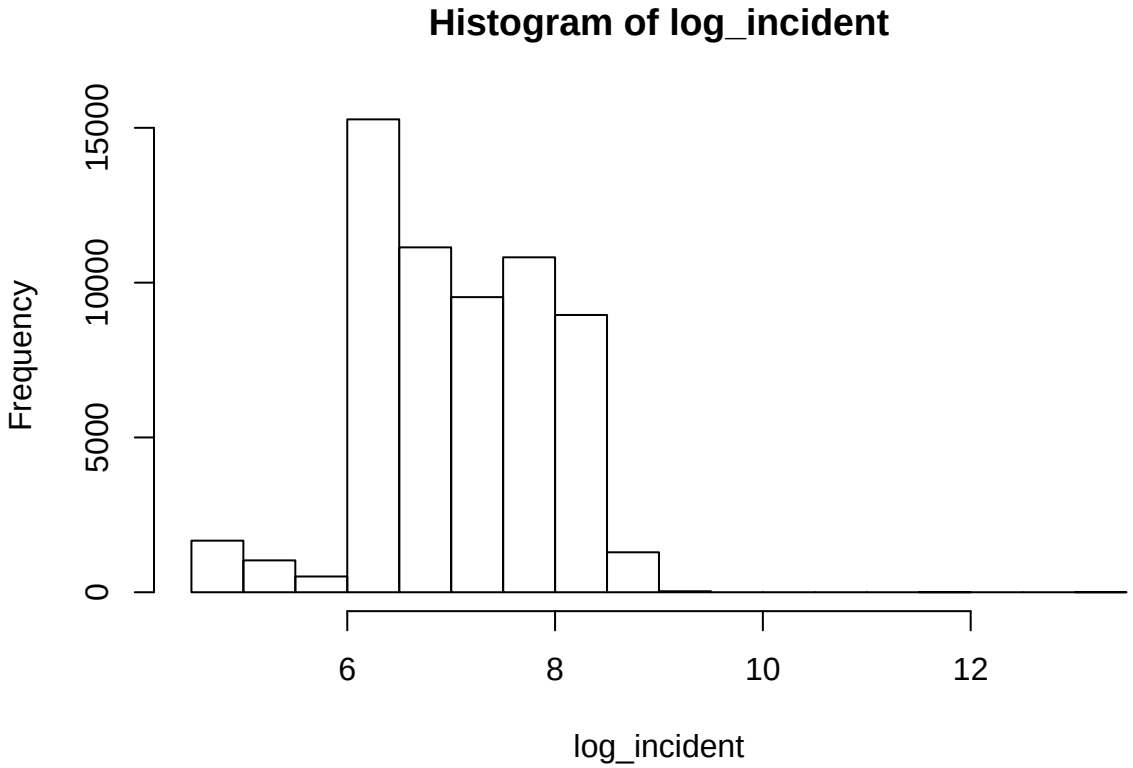


Figure 5: Log-transformed Incident PAR

Discussion of Box Plots

The boxplots in Figure 3[c] and 3[d] display different plots attempting to analyze the relationship between height of the sensor on the tree and temperature, relative humidity, incident PAR, and reflected PAR. Let's first consider the PAR readings of Figure 3[c]. For both PAR readings, we see a spatial trend in both the mass of the distribution. As you get higher on the tree, there is a much larger IQR of PAR readings, and more high-value outliers. This implies somewhat of a direct relationship between height and amount of light received; lower-level sensors receive less light on average than higher-level sensors. However, though the mass of the distribution at each height moves closer to zero, we see that the outliers in the incident PAR sensors still possess the full range of light readings at every height. On the other hand, as height goes to zero for the reflected PAR sensors, the outliers do not possess the same full range of readings present at the taller heights. Both temperature and humidity have a full range of values at every height, with similarly sized distributions and means. This implies that there is much less of a relationship between height and temperature and height and humidity than height and PAR. Figure 3[d] helps draw out the relationship between height and temperature. The plot for temperature shows that the lower heights are colder than average more often than the higher heights. For humidity it is shown that the higher heights are less humid than average more often than the lower heights. There seems to be a stronger relationship between humidity and height than temperature and height. This is due to the fact that there seems to be less outliers and tighter whiskers in the plot for relative humidity than for temperature. For the PAR readings, figure 3[d] brings to light the strong relationship between height and PAR value. But it also shows that there is high variance in the incident PAR values since all heights once again possess almost the full set of possible sensor readings. This is not the case in reflected PAR since the lower heights possess much smaller ranges of values than the higher heights. Also, it shows that there is not much relationship between height and reflected PAR, except for the tallest heights. These findings are all supported in the plots in Figure 4. Looking at the plot for temperature, it shows that there is not a strong relationship between temperature and height but it is clear that the top of the tree is hotter than bottom of the tree. This showcases a clear temperature gradient. The same can be said for relative humidity. Figure 4 illustrates a clear humidity gradient present across the

height of the tree that is also illustrated in the humidity plot in Figure 3[d]. Also, it is clear that humidity varies much more across the height of the tree than present since the range of values for relative humidity is much greater than the range for temperature in Figure 4. The plots in Figure 4 for PAR also support the findings in Figure 3[c] and Figure 3[c]. For incident PAR, there is a clear positive gradient across the height of the tree in Figure 4 and the large variation in incident PAR values that was shown in Figure 3[c] is also present. The plot in Figure 4 for reflected PAR shows the much lower variation across lower heights we found in Figure 3[c] and Figure 3[d]. Additionally, the plot in Figure 4 shows the positive gradient that was also displayed in the plots in Figures 3[c] and 3[d]. Therefore, I believe that Figures 3[c] and 3[d] convey the right messages and are successful in their deployment.

Suggestions for Improving Plots

The first two plots in Figure 4 serve to show how temperature and humidity vary throughout the course of a day and across the height of a tree respectively. However the plots are incongruous because the plots comparing temperature and humidity with time showcase the values for every height across every time value. However, the plots comparing temperature and humidity with height show the average sensor reading across all time steps for each height value. A good way to improve the readability of the first two plots in Figure 4 is to edit the time plots to show the average temperature and humidity readings across all heights at each time step. This would increase the readability of those two graphs while preserving much of the important information. The statisticians could directly give the reader the variation in temperature at different time steps if they would like to preserve that information.

Visualizing Network and Log Data

Yes, it is possible to generate a better visualization to highlight the difference between network and log data. Putting the data for both the network and log data in one graph would allow for a visualization that would better highlight the difference between network and log data. For example, consider a double box plot with data on the x-axis and average yield on the y-axis. With a plot like this, one can directly compare the average yield on any given day of network data vs the average yield on any given day of log data. Also, another plot that can be made in order to highlight the difference network and log data is a plot of yield percentage on the x-axis and height on the y-axis. With these two axis, two points can be plotted at every node height- for network data and log data respectively- to allow for a direct comparison of network and log data yield at every height. Pretty much, this plot is a combination of the third plot of Figure 7[a] and the third plot of Figure 7[b]. These two described plots would generate a better visualization of the differences between network and log data yield.