1. Data Collection

Project 1 Redwood Data Report

a. The study, “A Macroscope in the Redwoods”, attempted to capture spatial and temporal changes surrounding a coastal redwood tree through observations of its microclimate. There are limitations to traditional methods of collection, but through the use of wireless sensor networks, it was possible to collect enough data to perceive complex interactions such as the variation in humidity and temperature in the varying heights of a redwood tree. It was believed that the top of trees could experience a wide variety of temperature, humidity, and light while the bottom of trees were typically cooler, moister, and shadier. The goal of the study was to validate the substantial variation in temperature, humidity, and light that a single tree has over space. The study was forty-four days long in Sonoma, California and essentially documented “a month in the life of a redwood tree”. The data collected in this project was able to show the temporal and spatial trends in the microclimate of a redwood tree such as differences in light down the tree as light is absorbed by the canopy of leaves, lower sensors being cooler than the average temperature due to insulation of the canopy, and the quick and drastic change in humidity just prior to sunrise. The way in which data was collected was also very impactful and much could be learned from it. The design and testing of the instruments were important to make sure that accurate measurements would be made. The data collected was either through the sensor network or through the data log in the nodes, the use of both sources was important as researchers wanted to learn more about how well the wireless network performed and also considered possible mishaps in the methods such as lack of memory storage or network issues to ensure quality data was obtained, so any conclusions made could be trusted. There were some differences between the data collected by the sources, so this stressed the importance of having both a wireless network and also a flash log.

b. As the main contribution to the data was being recorded on a seventy meter tall redwood tree, a strategic method for data collection was needed for accurate readings. The duration of the data recorded spanned the course of approximately forty-four days as the first and last recorded pieces of data were on April 27th at 5:10 pm and June 10th, 2004 at 2:00 pm, respectively. The data collected were done by various wireless sensors that were strategically placed throughout the tree. To take account of the weather conditions that could potentially harm the wireless sensors throughout the duration of a month, the sensors were specially packages in a way that would not be impactful to the readings. The sensors were placed along a vertical range that spanned from fifteen meters to seventy meters with a vertical spacing of approximately two meters between each of the sensors. The reasoning for starting the placements at fifteen meters and not at the base of the true was due to the fact that microclimate variations occur at a higher level when the altitude of the tree is greater, so data collected at higher altitudes would have more significance to the data. Horizontal placement also comes into effect as the sensors were placed from a horizontal range of a tenth of a meter to a meter away from the main base of the tree. They were placed near the base of the tree as direct variation was desired and closer sensors lead to more accurate readings. The sensor outputs were also dependent on the sensor orientation so the sensors needed to be placed so that they were leveled. The sensors also preferably had adequate airflow and a shaded area as they did not want the readings to absorb solar heat to throw off the humidity and temperature readings. The data recorded from these sensors were also done in a way such that data was recorded for every five minute throughout the duration of experiment. The data was collected in two various was that either incorporated a sensor network or through a data logger. An application that was used for the purpose of data collection was the TASK software. The TASK software would wake up the sensor for about four seconds to record its variables for

every five minute interval. These characteristics were able to be recorded through the selection of various wireless sensors. In the scenario where the networks did not work, there was an incorporated backup system which was a local data logging system. The data logger would store readings before they were passed through a multi-hop routing layer and was stored on a 512 kB flash chip that was incorporated with every sensor package. The decision on what features were important enough to record was narrowed down to selecting biological features that could give lead to more insight. The climate variables that were then chosen were temperature, humidity, and light levels. These were chosen as they give more insight on the information about the forest transpiration and the carbon balance that is present.

2. Data Cleaning

a. Most of the variables seemed to have relatively similar distributions when comparing across the two data sets from the flash logs and the wireless network, but the variable epoch had very different distributions seen in *Figure 1* and *Figure 2*. The values of epoch had different ranges in the two datasets the flash log data had values from 0 and onward while data from the wireless network only had values after about 2500. The data in the wireless network was also somewhat uniform while it was skewed right in the flash log data. This difference in distributions could have been attributed to the flash log data being collected from before the nodes were set on site and the lack of data from large values of epoch since epoch is related to time and some of the nodes died towards the end of the study.
A variable that seemed inconsistent across the two data sets was the voltage. Values for voltage in the in the wireless network dataset were above 200 while values in the flash log data set ranged from about 0 to 3. Reading the article, it seems that the voltage variable in the wireless network data was measured in an ADC reading and the voltage variable in the flash log data was measure in volts. We decided to convert the voltage column in the wireless network data into volts to match with the flash log data. Looking up the conversion, it followed that 1023 ADC reading was 5 volts, so to compute the conversion we multiplied each voltage value in the wireless network data by 5 and then divided by 1023 to convert to volts. Even after the conversion, the voltage variable distribution in the two datasets still differ as most of the voltage value in the wireless network data was around 1 and the value in the flash log data was around 2.5 and 3 seen in *Figure 3* and *Figure 4*.

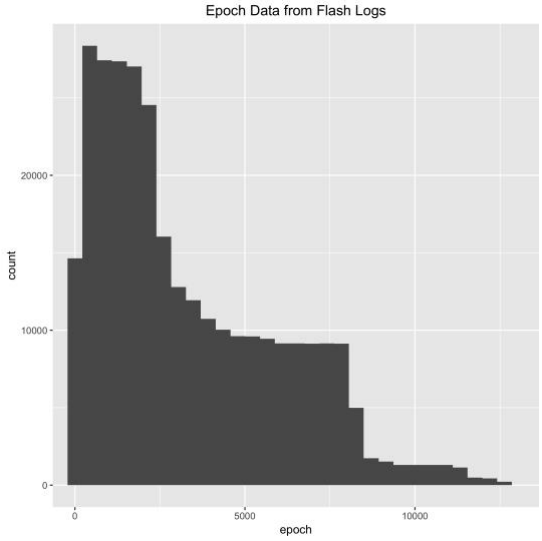


Figure 1

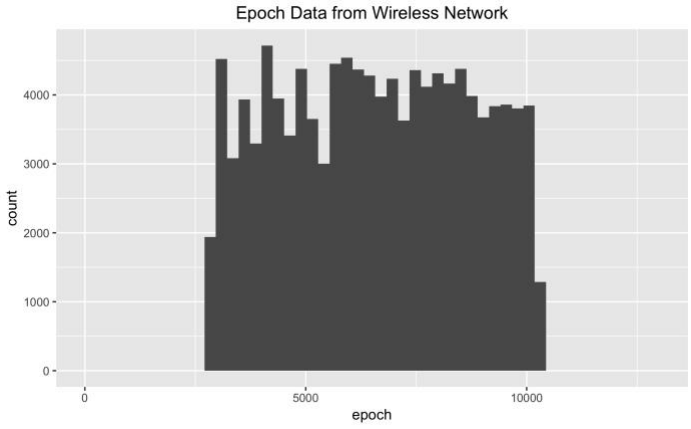


Figure 2

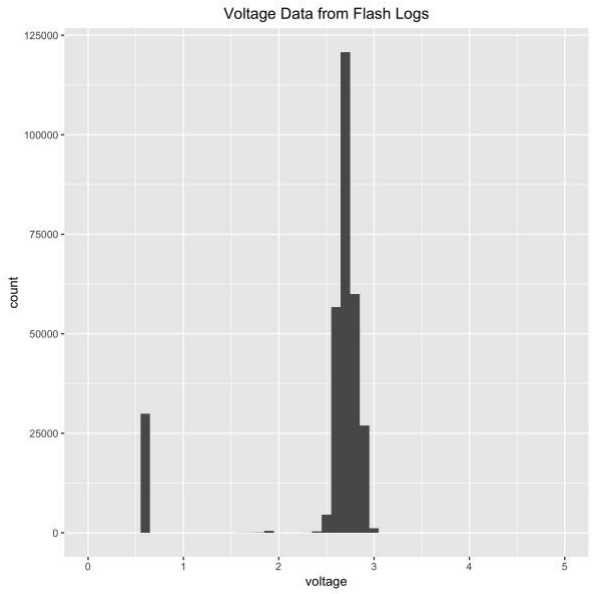


Figure 3

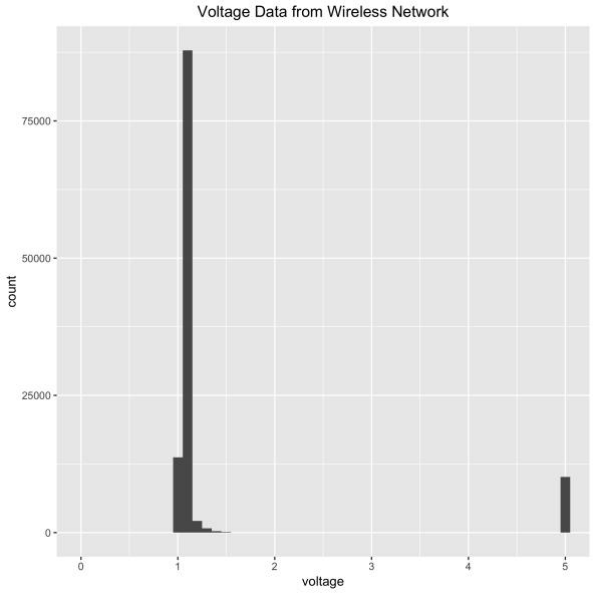


Figure 4

b. The number of missing observations in the flash log dataset is 8,270 which relates to missing 41,350 individual measurements while the number of missing observations in the wireless network dataset is 4,262 relating to missing 21,310 individual measurements. The corresponding dates where the flash log data set is missing measurements is November 10, 2004 according to the result_time column, but using epoch the dates are April 30, 2004. The dates in the wireless network dataset where there are missing measurements is in the time period between May 7 and May 13, 2004.

c. The number of variables in the new data frame after incorporating location data is 15. The variables are as follows: nodeid, result_time, epoch, parent, voltage, depth, humidity, humid_temp, humid_adj, hamatop, hamabot, Height, Direc, Dist, and Tree.

d. Humidity seems to have at least one large negative outlier and since it is measured on a percent scale, so it would make sense that the values should only be between 0 and 100, so that is why we decided to remove any data that lies outside of that range. This resulted in the removal of 22,279 data points which seemed okay considering there were over 400,000 data points. Temperature or humid_temp seems to have outliers both greater than or less than the majority of the data. Looking at the default histogram the outliers seemed to account for a very insignificant proportion of the data while the majority seemed relatively close to the mean, so we decided that removing outliers beyond 2 standard deviations would remove anything very out of the ordinary that may have been a mistake. This removed 2,224 observations. Incident PAR or hamatop is a measurement of light with a couple of very large positive outliers. We decided to remove any data points that had values of incident PAR less than 0 since it was a measure of light and there is no negative values for that and then decided to remove any values above 2500 by looking at the histogram which resulted in values that might have seemed like outliers, but there was a small cluster of them which we made the decision to keep in the event that they represented some sort of phenomenon that the other points did not cover. Our cutoffs resulted in the removal of 60 observations. Reflected PAR or hamabot is another measurement of light with large positive outliers. We decided to remove any data points outside of the range of 0 and 2.5 standard deviations above the mean since there should not be any negative values for this measurement (there were none) and there seemed to be a couple of clusters of large positive values as well similar to

incident PAR that we felt might possibly be representative of something going on and did not want to just remove them. This decision resulted in the removal of 10,282 data points.

3. Data Exploration

a.

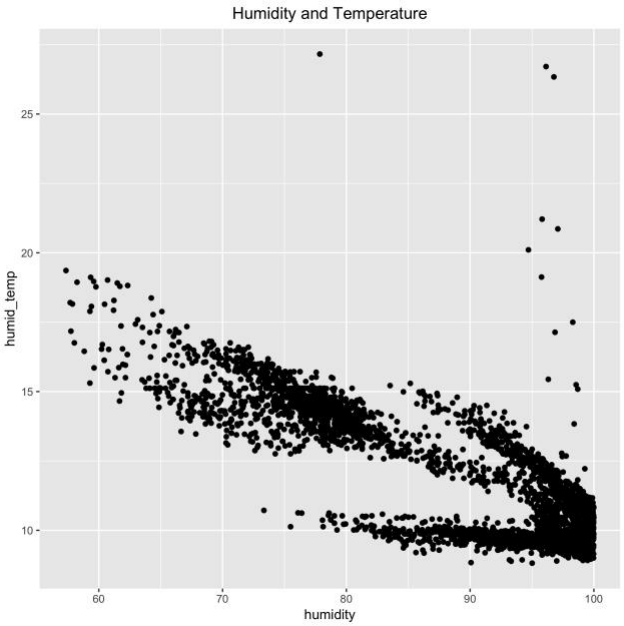


Figure 5

We decided to analyze a subset of our data for the date of May 8th to plot some of the variables against each other and to look for any relationships. Humidity and temperature had a somewhat negative relationship with one another as it seemed that as values of humidity lowered, so do temperature see *Figure 5*. This was also apparent in plotting humidity and temperature against epoch which is a measure of the date or time in this case as the two scatterplots, *Figure 6* and *Figure* 7, look somewhat like reflections of one another. Humidity seemed to be high for a significant portion of the day and then drastically decreased and temperature seemed to be low for that significant portion with high humidity and then rose in temperature similar to when humidity drastically decreased.

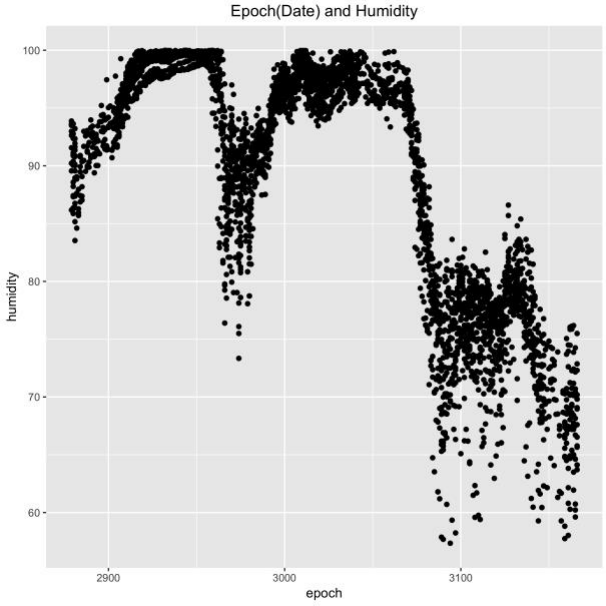


Figure 6

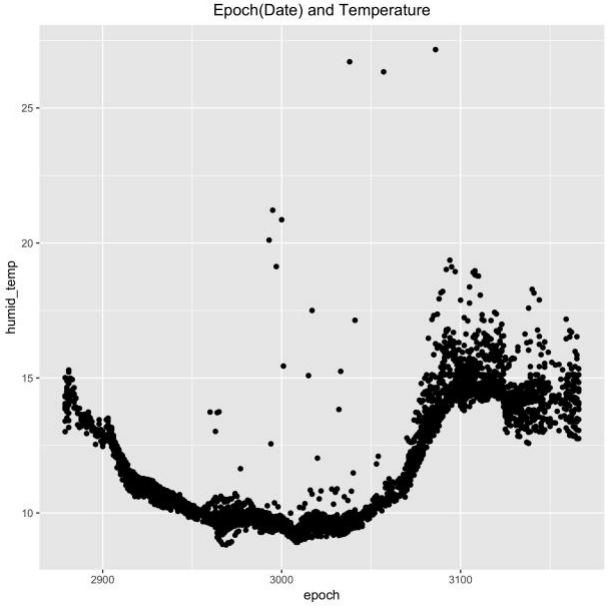


Figure 7

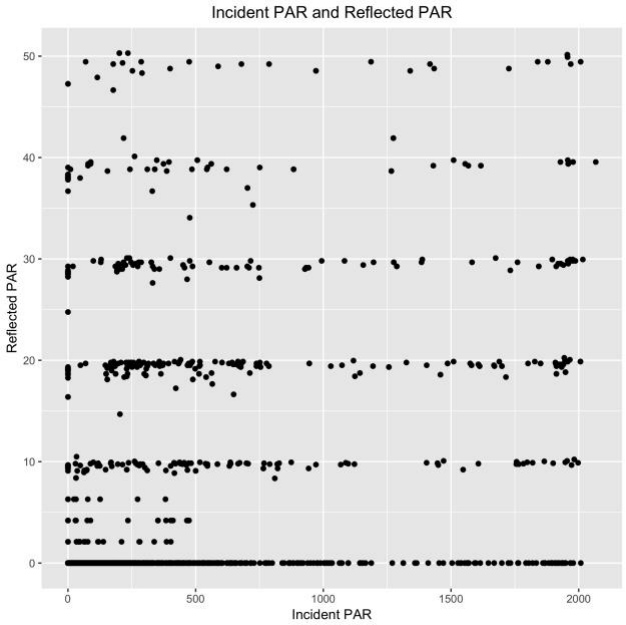


Figure 8

Incident and Reflected PAR are both measures of light and we found the scatter plot of these two variables, *Figure 8*, to be interesting since there does not seem to be any sort of linear relationship or any relationship at all really between the two which was surprising given what they were measuring. Although with reflected PAR, light had a more difficult way to travel, we would assume no relationship for values of 0 or possibly small values, but for strong readings indicating higher levels of sunlight, it seems odd that these do not directly correlate with higher levels of incident PAR.

b. Incident PAR seems to be associated with quite a few of the predictors. The relationship between incident PAR and depth is positive, so when there is smaller measure of depth there seems to be more variation in light typically with higher values of incident PAR in general. Unlike depth, height has a positive relationship with incident PAR with higher values of height leading to higher readings of incident PAR. There also seems to be a relationship between humidity and incident PAR, when humidity levels are higher, values of incident PAR or light are lower; the same follows for humid_adj or relative humidity. Temperature is also associated with incident PAR with a relatively positive relationship where lower temperatures are aligned with less light or measures of incident PAR. There also seems to be some relationship between the direction and incident PAR where certain directions have higher measures and variation of incident PAR.

c. Over the course of a single day, in this case we chose May 8th, temperature seemed to follow a trend regardless of height, except for a couple of points at high heights that seemed to experience much higher temperatures somewhere in the middle of the day (*Figure 9*). Humidity also seems to follow a trend over the day, but higher heights typically have lower humidity (*Figure 10*). Although in the couple of hours prior to midnight near May 9th, the humidity seemed to drop for nodes at smaller heights. Incident PAR stayed relatively low until midday where nodes at higher heights experienced higher values of incident PAR they also experienced slightly higher values earlier on in the day nearer to May 7th (*Figure 11*). However, a small cluster of values at very small heights did not follow these trends and had fairly consistent values of incident PAR over the entire course of the day. Similar to incident PAR, reflected PAR did not experience very high values until midday and later (*Figure 12*). There was more variability in the data for each hour, but for the most part, nodes at higher heights experienced higher values of reflected PAR.

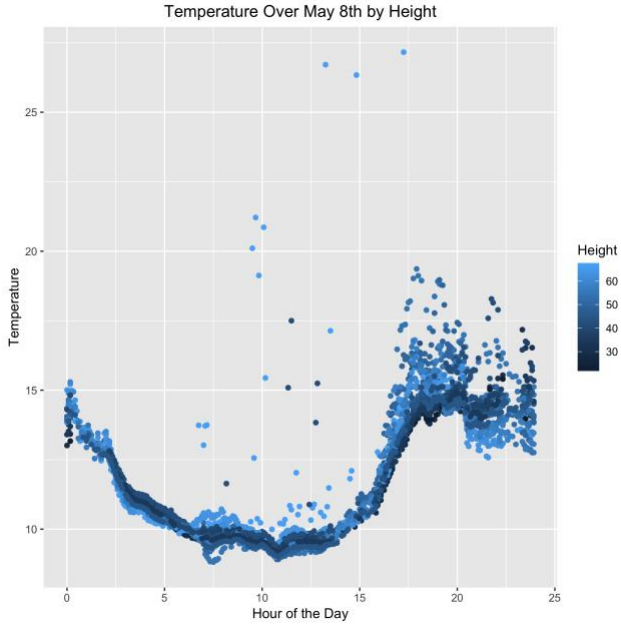


Figure 9

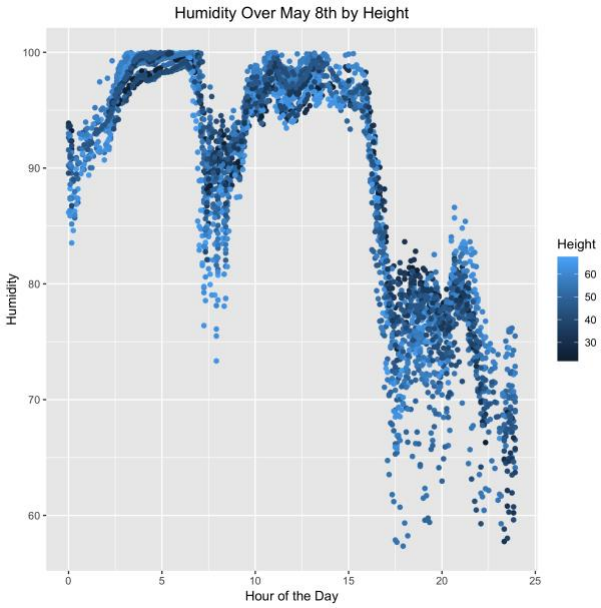


Figure 10

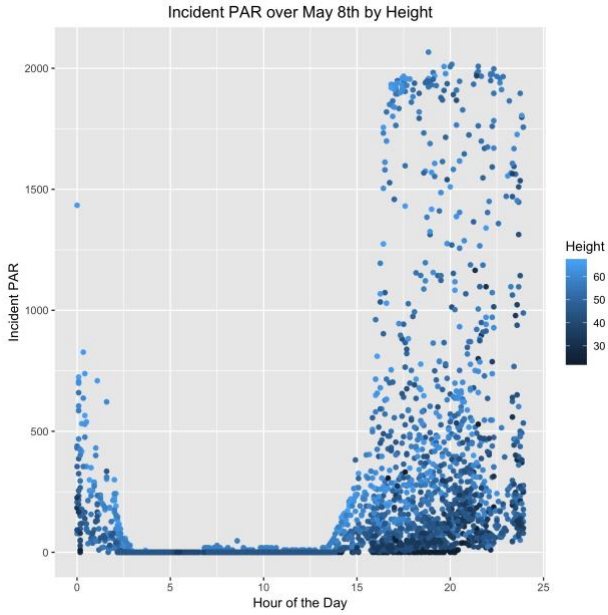


Figure 11

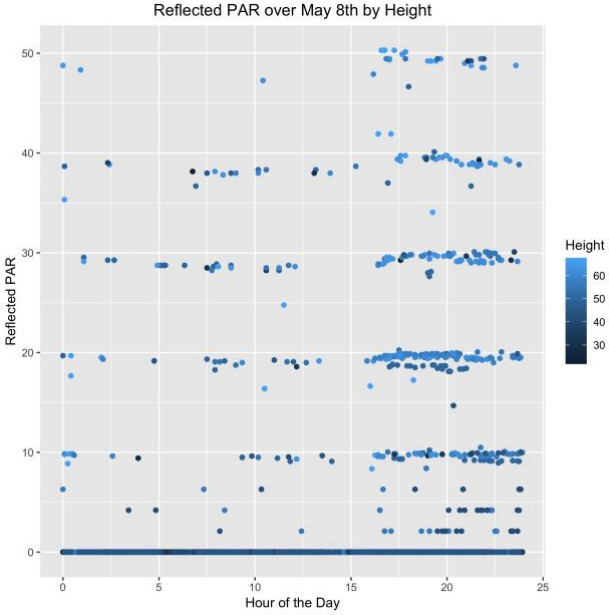
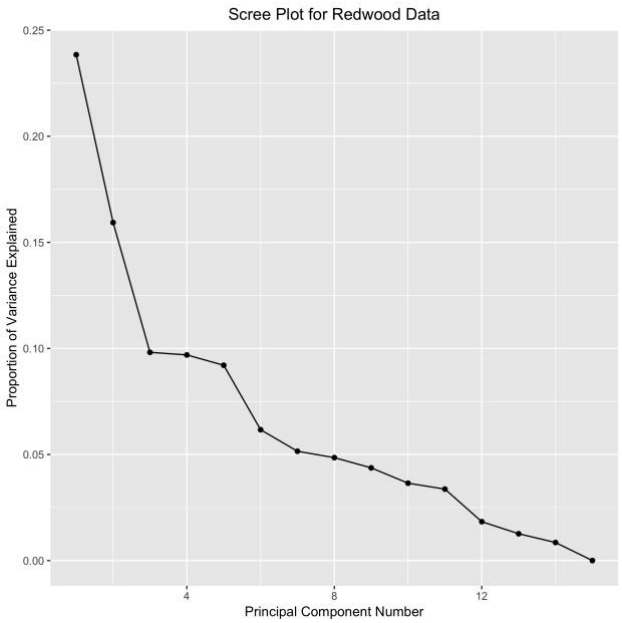


Figure 12

d.



This data cannot be approximated by some low dimensional representation since the first two or even first three principal components only account for about half of the data's variation according to the scree plot see *Figure 13*. If we were to use a simple scatterplot of the first two components, we would lose a lot of the details in the data since so little of the variation is accounted for.

Figure 13

4. Interesting Findings

a. After running principal component analysis, we used the first and second PCs to run k-means on the data for two clusters in *Figure 14*. The principal components seem to have a split as there are two distinct lines that appear when plotting and the k-means algorithm does an okay job of distinguishing between the two although it does seem to do rather poorly for smaller values of PC2. This is rather odd due to the fact that the algorithm searches for points to two different means in the data to assign what observations belong to what cluster.

b. We also decided to run a gaussian mixture model algorithm on the data as well to try to distinguish between these two line "clusters" that we see when plotting the principal components in *Figure 15*. GMM did a better job than k-means it seems as it separated the "clusters" relative to their trends on the principal components. Similar to k-means it seemed to misclassify values more often where the second principal component was lower. The increased accuracy in using GMM over k-means was to be expected since the data did not particular have the cloud like clusters that work well for k-means. The only advantage would be the efficiency of k-means in this case since the data is quite large.

c. In analyzes the differences between the data collected from the wireless network and from the flash logs and since the plot of the first two principal components seemed to show two line "clusters", it brought up the idea that maybe the two "clusters" corresponded to the two different datasets. We did a rough split of the data for those whose first and second principal components were nonnegative to match the first line "cluster" and data whose first and second principal components were negative to match the second line "cluster" and drawing histograms for the epoch variable (*Figures 16, 17*) since this showed some of the difference between the two datasets. This did not completely support our idea since both "clusters" had epoch values from 0 onwards, but for one of the "clusters" it had extremely few values above 3,750, so it follows that one of the clusters was only representative of a specific time period during the study while the other one covered the entire time.

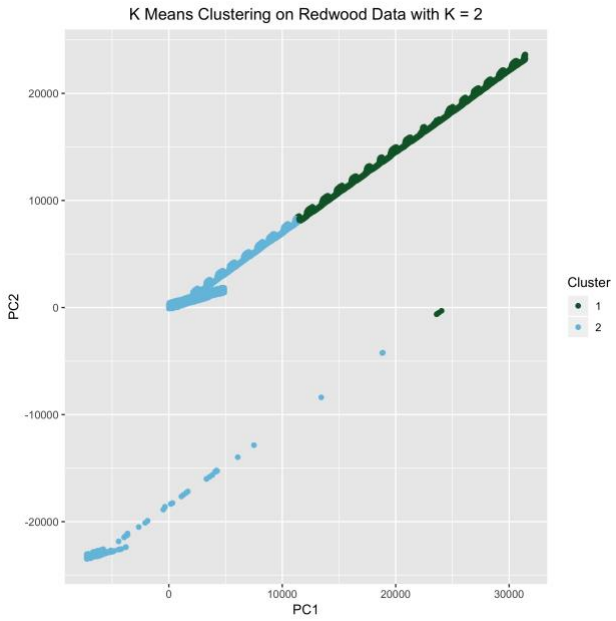


Figure 14

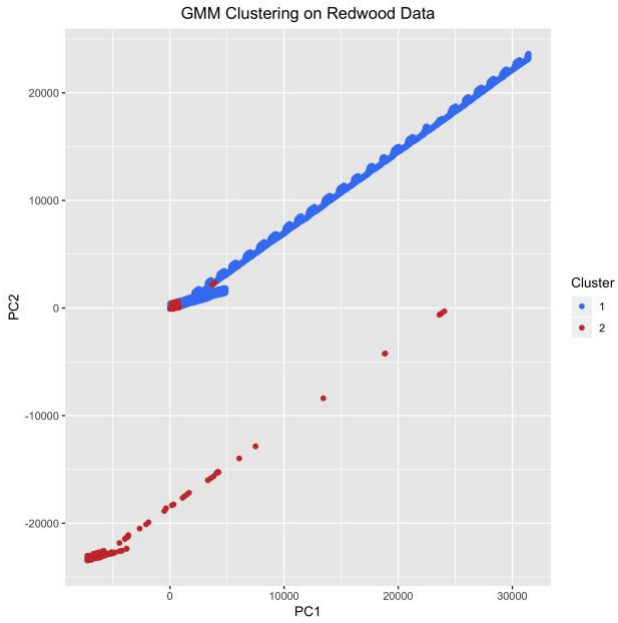


Figure 15

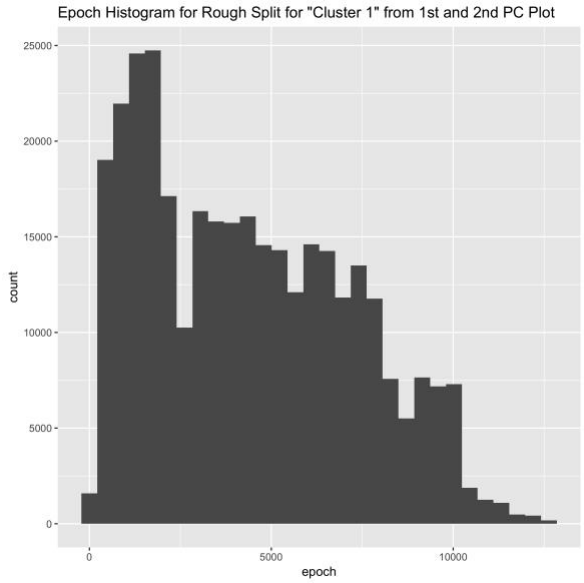


Figure 16

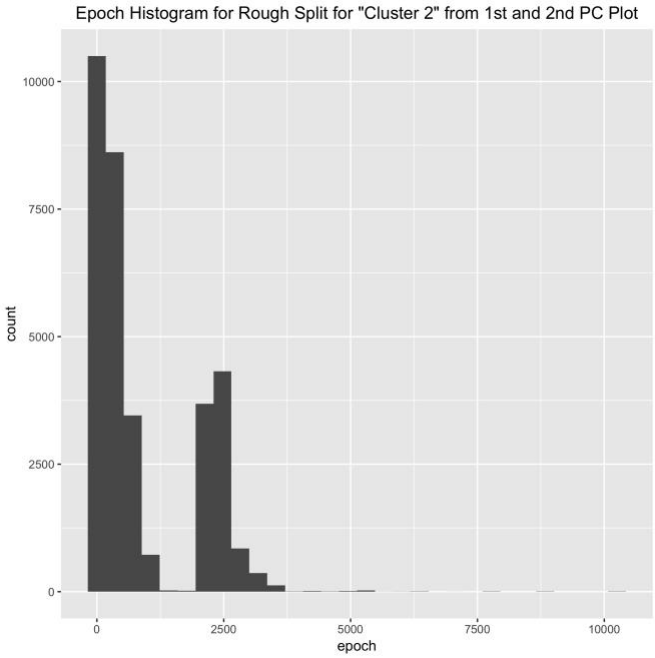


Figure 17

5. Graph Critique in the Paper

a. The range of values for both have extremely long tails, so to better visualize the data we performed a log transformation of the data. Since some of the values of both incident PAR, hamatop, and reflected PAR, hamabot, are 0, we also shifted the original values by 1 to account for this.

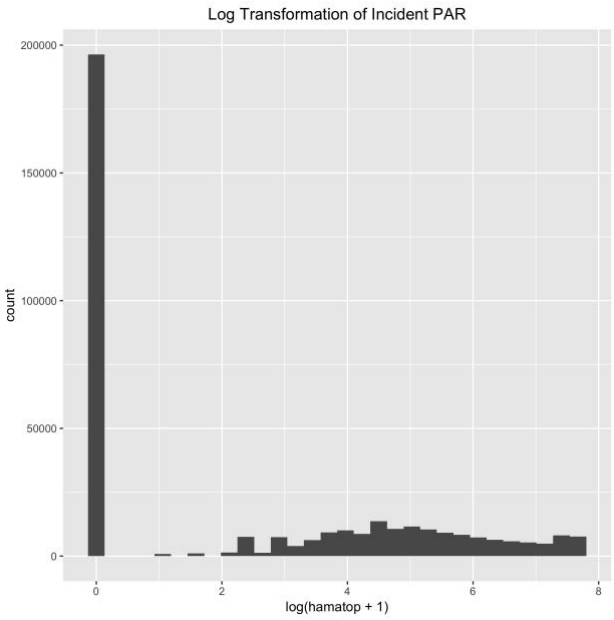


Figure 18

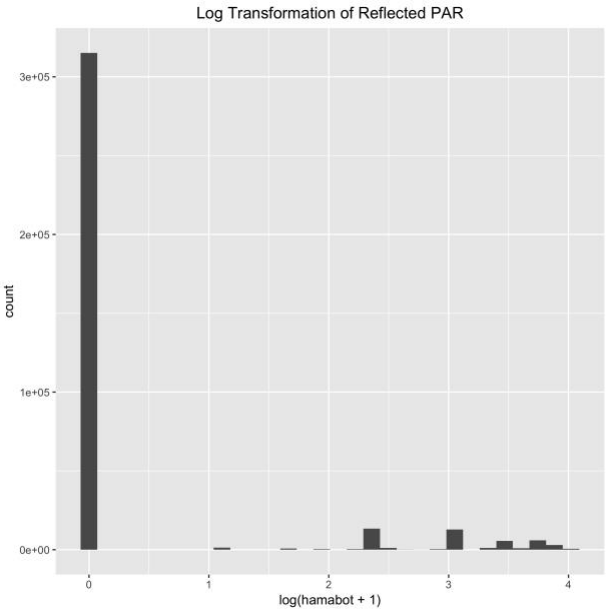


Figure 19

b. The boxplots of figure 3(c) and figure 3(d) in the article display a summary of the characteristics and distribution of each of the characteristics relative to each of the heights. The distribution is described in terms of the median and its interquartile range of each height over the entire course of the day. The plots may not convey the right message because we are trying to better understand the microclimate of the tree which is the difference in temperature and humidity across the height of the tree at a given time. The boxplots don't really take into account the differences in time, but rather gave an overview of the variation in temperature, relative humidity, incident PAR, and reflected PAR over the different heights. From a data visualization standpoint, it is also hard to read and interpret for most readers although the detail on the distribution may be helpful for close analysis. The graphs are also hard to read due to their size and the fact that there is a distribution for each height from fifteen to seventy meters. We thought that despite losing some information such as the median at each height, visualizing overall trends in the data according to time such as the prevalent difference in height for similar values of temperature and humidity, a scatter plot with height as a color component does a much better job of telling the story of the changes in the microclimate of the tree. There are also many points to plot and to deal with overplotting, we increased the transparency and reduced the size of the points to increase the visibility of the data (*Figures 20,21*).

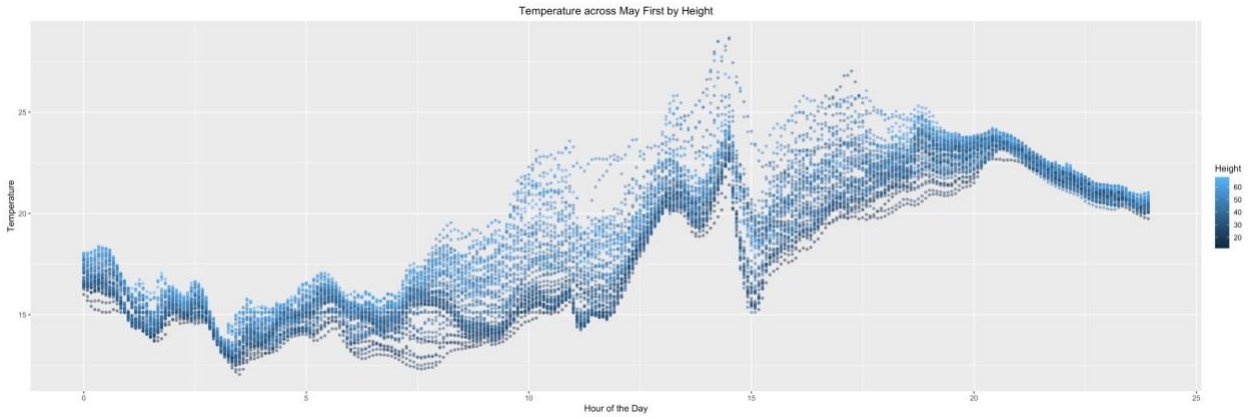


Figure 20

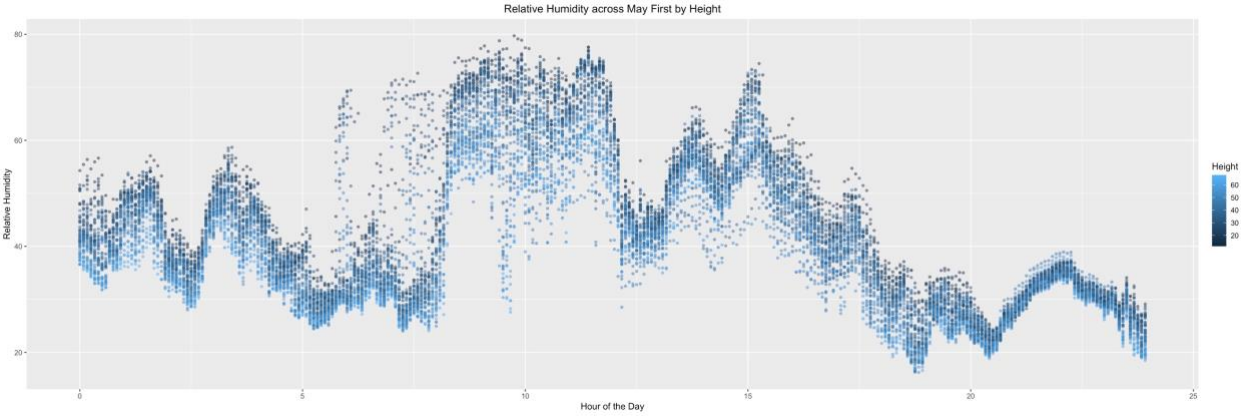


Figure 21

c. The biggest problems with the first two graphs of Figure 4 is that the color choices representing the heights do not really say much and since there are so many different heights it makes it nearly impossible to distinguish between the different heights and even more difficult to determine which are high height values and which are low. This lack of distinction between height values does not allow for height to be seen as an important factor when it is. In order to remedy this, we decided to plot height using colors on a gradient scale where darker colors indicated lower heights and lighter colors corresponded to higher heights since there were so many heights to account for and what really matters are the relative values to one another. The choice of using a line plot also makes it difficult to see since the added concatenation of the points just adds more clutter to the graph, so simply plotting the points and changing the transparency and size helps with the possible plotting of so many points. Given that there is so much data to account for and we felt averaging the value may be too much of a reduction in the data, the graphs from part b seemed the best.

d. Figure 7 in the article is a data visualization of the differences between the two recordings of the log and network readings. A way to improve the data visualization could be to combine the graphs and distinguish the two through different color representations to make comparison easier for the third and fourth plots. The use of bars is also somewhat cluttered, using thick lines instead may reduce the clutter while conveying the same message. The use of dots in the third plot make it difficult to determine values, using lines drawn up to the percentage point might make it easier to read as well. The third and fourth plots in the figure with percent vs node height and day may also benefit from a switch in the axis as the y axis is difficult to read with all the many heights. Making the individual plots easier to read would help in analyzing the differences between the two datasets as the information is more digestible.