STA521 Project1: Redwood Data Report

Shuo Wang(2710299), Tianhao Li(2710173)

September 23, 2021

1 Data collection

1.1 Summary of paper

The main purpose of the study is to use deployed sensor network and the state of the art in the measurement and analysis techniques to obtain and reveal trends and macroscopic views of previously unrecorded phenomena, which biologists would like to observe but were previously-unobtainable.

In the study, the data was collected from 70-meter tall redwood trees in a study area called the Grove of the Old Trees, in Sonoma California, with a time period of 44 days. To be specific, the data was collected in one month during early summer, sampling once every 5 minutes. The sensors were deployed from 15 meters from ground level to 70 meters from ground level, with roughly 2-meters spacing between nodes. Additionally, most data was collected on the west side of the tree, within a 0.1 to 1 meter range.

The main conclusion of this paper is that the deployment and analysis of the sensors network yielded valid temporal and spatial gradients which captured the complex environmental dynamics of the microclimate and can be used to build quantitative models. Offering the potential to advance the state of science by enabling dense temporal and spatial monitoring of large volumes, this network and data allow scientists to validate biological theories which involve things they cannot measure today, like models on the sap flow rate and large-scale process of carbon and water exchange.

1.2 Summary of data collection

The data is collected in a very organized way, with the latest TinyOS and TASK software for the node operating system, networking stack, and data collection framework. The sensors used are integrated with an existing wireless sensor node platform, and first are checked by two calibration phases: roof and chamber, to separately examine performance for PAR and for temperature and humidity.

As for the employment in the tree, every four sensors are packaged in one mote, with a total of 33 motes. The motes are mainly deployed on the west side of the tree. Regarding the vertical distance, they are 15m from ground level to 70m from ground level, with roughly a 2-meter spacing between nodes. Regarding the radial distance, they are 0.1-1.0m from the trunk.Several other nodes are employed outside of this region.As for the duration, data was collected during 44 days in the early summer, sampled once every 5 minutes.

The traditional climate variables are of interest, including temperature, relative humidity, and photosynthetically active radiation (PAR). PAR measurements include both incident (direct) and reflected (ambient) levels.

As for the data organization, each column represents a particular sensor and each row represents readings taken at a particular time. Besides, node ID, sample number, and sample reception time are also represented as columns in this table.

The data retrieved over the wireless network is stored in "sonoma-data-net.csv". In this case, the data is recorded every 5 minutes and is transferred back to the base station immediately. The data retrieved from the flash logs after the deployment is stored in "sonoma-data-log.csv". In this case,

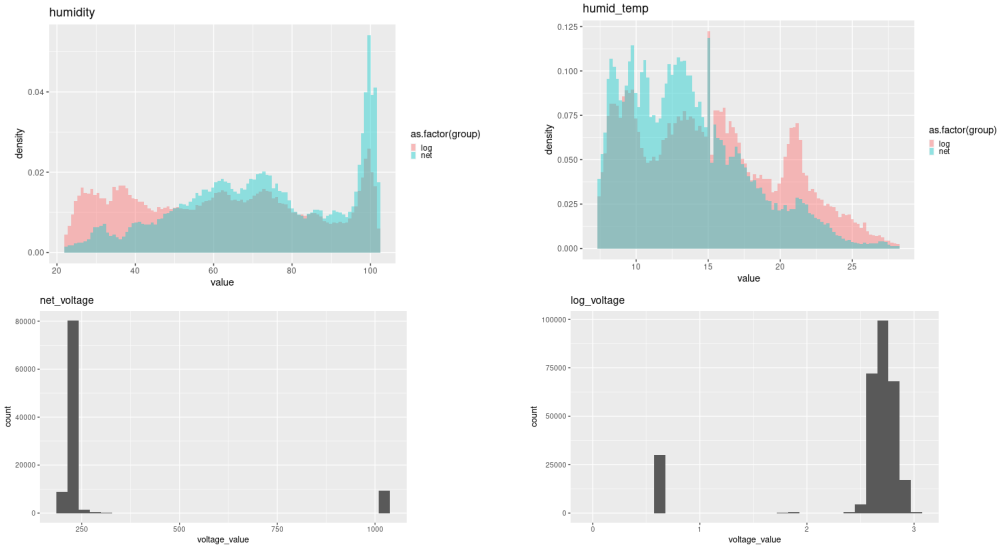
the data is stored in a local data logging system and is retrieved only after the whole deployment process. The logger recorded every reading taken by every query and stopped recording once the 512 kB flash chip was full.

2 Data Cleaning

2.1 Check variables consistency

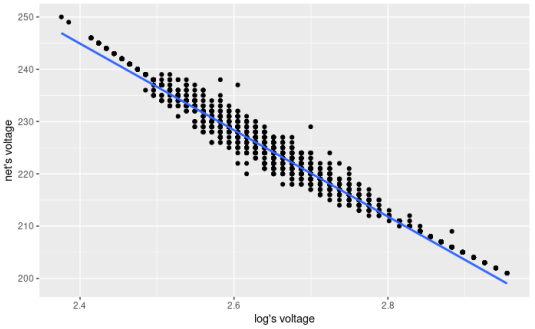
In this part, we first load the data from original files. Here we noticed that the date time in the csv files are problematic, while the dates and time in the sonoma-dates file provides the true dates and time and the corresponding relationship. So first of all, we convert the dates file to provide the accurate dates and time.

After doing the time conversion, we looked at the data again and found some duplicate value in the log and net data which have the same epoch and nodeid and many other attributes while only result time or few other attributes are different, so we decided to deal with them. Then we were able to check the consistency of the variables.



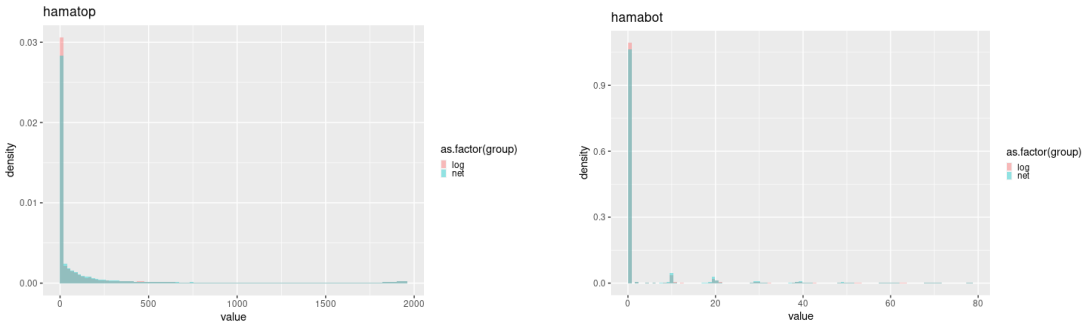
From the plots we can see that hamatop and hamabot which represent photosynthetically active solar radiation are generally consistent, but network data of humid_adj and humidity concentrate more on the larger values and are more left-skewed. Network data of humid_temp concentrate more on the smaller values and are more right-skewed. Here we use the histogram of densities because the counting histogram is not appropriate. The log and net data sets have different amounts of data, which will lead to a disproportionate plot. Another trick we used is specifying the same interval of two datasets to get the same bin which enables us to draw the two plots together.

Here we noticed that the voltage data in two dataset is very different, but it seems that there is a linear relationship between them, so we tried to apply the linear regression model to convert the data. We filtered out node 134,135,141,145 whose readings are all above 1000 voltage which are obviously outliers. Then a linear model was fitted and we obtained a sensible result with a 0.9961 R^2 and a very small p-value of less than



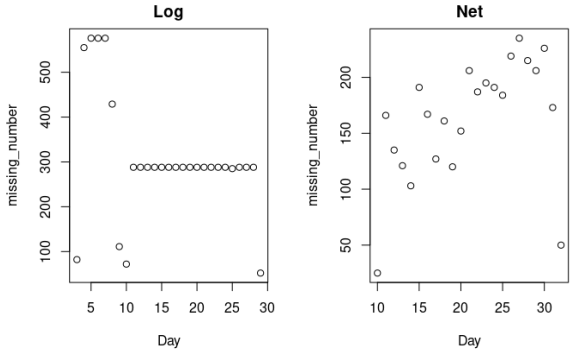
2.2e-16. We think this result is acceptable and could be used as the converting method. We used this linear model to convert the voltage readings in the net data set to a scale of 2-3 voltage. One thing to mention here is that the nodes we filtered out here show normal readings of humidity and temperature and hama values in section 2(e), so these points are kept in this project. In the end, the voltage readings in the net set were converted into the same range as the log set using this linear model.

Besides, we noticed that the ranges of hamabot and hamatop are different from the ranges in the paper. After doing some research, we found out that the unit of hamas in the paper is PPFD, while the unit in the data set is LUX. Thus, in order to convert the data unit to be consistent with the paper, we need to divide the hamas by 54. After the conversion, we can plot these plots again. The 2 plots show a high consistency in hama readings.



2.2 Missing data

Here we need to mention that after we incorporated the date and time information, we found that only dates in May and June existed. These two plots show the number of missing values on different days. Here we see a constant 288 missing data from 11th to 28th and a large number of missing data (455-576) in the Log dataset. In the Net dataset, the missing data is more sparsely distributed. The constant reading in the log data may suggest some systematic problem.



2.3 Incorporate location

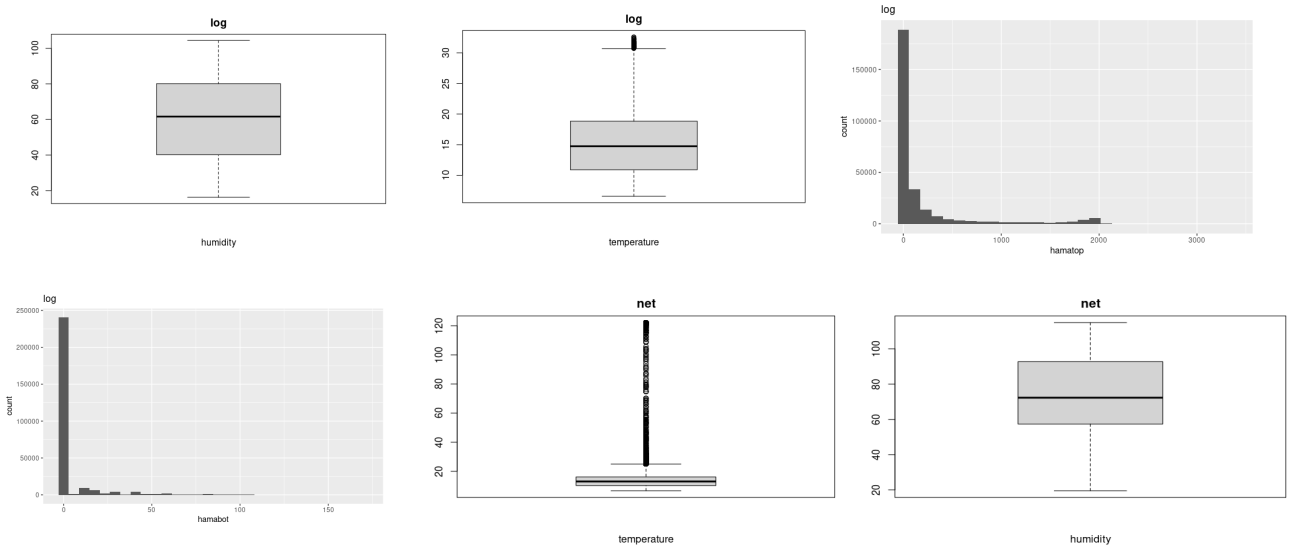
The location data is separate in another file mote-location-data.txt and we use the node ID as the index to pair these two data sets and to incorporate the location information. Here we noticed that node 65535,135 and 100 in the log and net data file do not have a corresponding ID number in the location file. By using the ncol() command, we saw that there are 16 variables in the data frames(both log and net) called "log_tl" and "net_tl". We would like to mention that the "epochDates" variable in our data frames consists of the week day and date and time.

2.4 Outliers

First we took a look at the log data set and removed the data with humidity less than 0 which is obviously wrong. Then we checked the humidity and temperature. By drawing 2 box plots, we can

see that actually there are no outliers in the humidity readings. This was confusing at first, since we expect humidity to be not greater than 100%. However, after searching the Internet we found out that if the humidity is a "relative humidity" then the humidity can exceed 100%. Besides, since there is no outlier in the box plot, we decide to keep all the humidity data at hand. As for the temperature, we can identify some outliers through a box plot, then we exclude the data where the temperature is greater than 31 degrees Celsius.

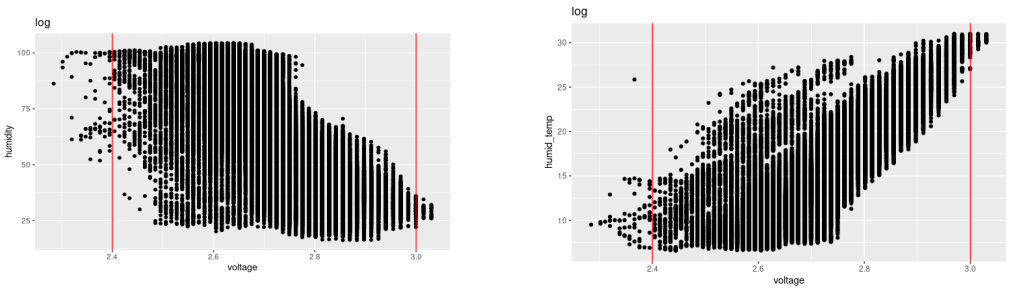
Finally we took a look at the PAR readings. Here we noticed that if we use a box plot to identify the outliers in hamatop and hamabot readings, then we simply throw away most of the hama data since there are many zeros or values close to zero. However, we would like to keep these values in order to keep some useful information. Thus we used a histogram to have a view of the data distribution and manually decided data under which quantile we would like to keep. When deciding how much hama data to keep, we found it hard to set up a consistent criteria for both data set to generate similar results with the original paper. We think maybe some expert knowledge was involved to set up the range of readings, so we used the same criteria here.

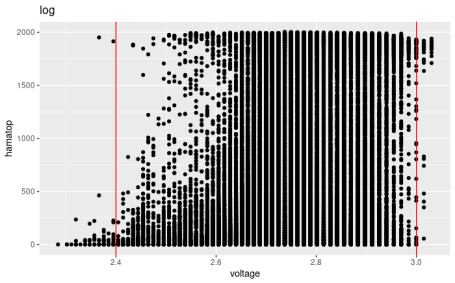


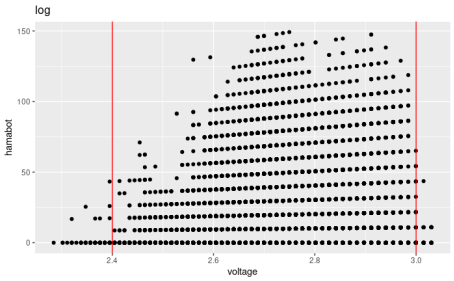
The operation on the net data set was quite similar to operation on log data set. We first used the box plot to filter the outliers in humidity and temperature data, then used the criteria in the original paper to filter the PAR(hama) data.

2.5 Other outliers

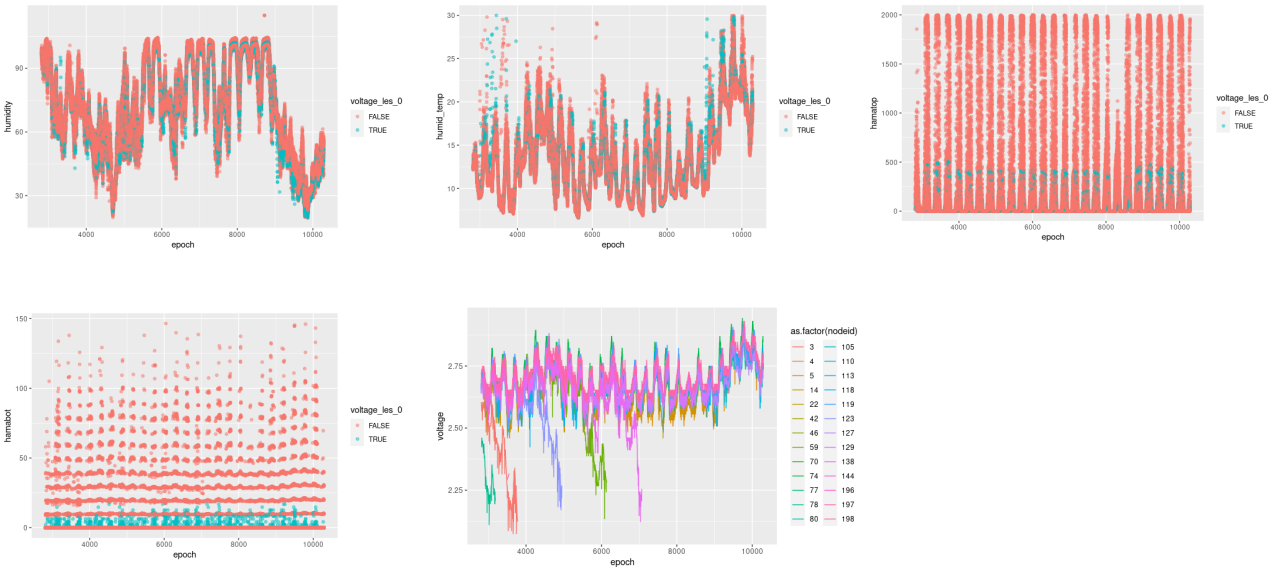
First we checked the log data set. We saw data with a voltage reading less than 2 could be outliers. By considering the overlap of 2 data set, we found that the intersection of two data sets consists node 134 141, 145, whose net voltage readings are also abnormal. However, below we are going to illustrate that their other readings are actually quite normal. By taking a look at the scatter plots of voltage versus humidity, humid_temp, hamatop and hamabot, we can see that the red lines represent the voltage value of 2.4v and 3v in the log data set. If we just cut off the data outside these lines, we would throw away some data which is actually not an outlier. For example in the voltage versus humidity plot, we are about to throw away some normal plots in the cluster and create a biased data set against high humidity and low humidity. This case is the same for the other 3 scatter plots. Thus, we would like to keep these points.







In the original paper, the author uses Figure 6 to illustrate that the voltage readings of 4 sensors are related to their abnormal temperature reading so that they removed those nodes data with voltage readings under 2.4V or above 3V. Despite the reason in the previous paragraph, in this section we found out that after filtering temperature and humidity, we still observed some deviation the the voltage plot against time (here we use epoch to simplify the presentation, they are basically the same). Now we consider the mutual points between net set and log set, the ones that have low reading in log set also have negative reading in the net set after conversion. However, from the plots here we can see that actually their other readings behave just like normal points.

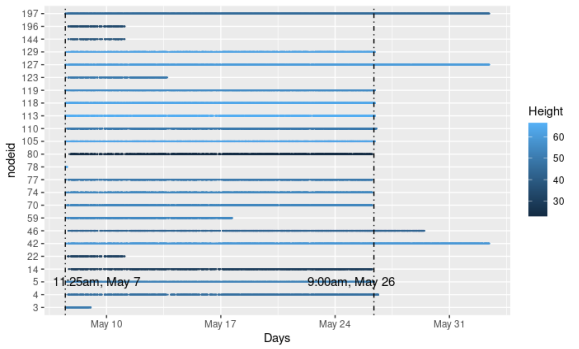


3 EDA

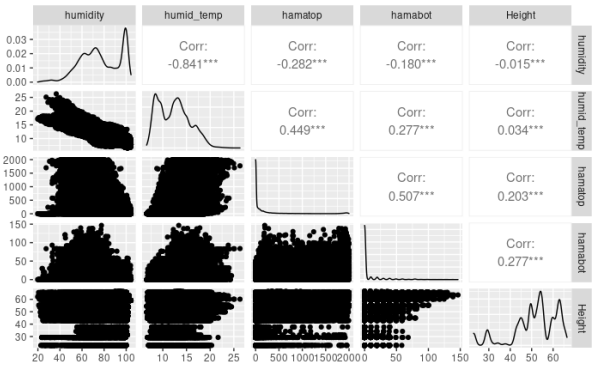
First of all, before any data exploration, we need to specify the data we are going to use. Here we chose to merge the log data set and net data set, since the combination of these two data set provides the most comprehensive data. As mentioned in part 2, we had converted the voltage scales. Here we use the appropriate voltage readings as the readings of the combination set. Along with combining the data, we excluded duplicated data and focused on the interior trees. The final combined data set accounts for 74% of the net data and 26% of the log data.

3.1 Scatter plots

We chose the time period from 11:25 a.m. May 7th to 9:00 a.a. May 26, because this period contains most of the available data and can show the cumulative effect across time.



These pairwise scatter plots seemed messy at first glance, so we also tried to plot the plots of one day.The pairwise scatter plots in one day are similar to these plots, the only difference is that they are sparser. Thus, we decided to stick with the long time period.



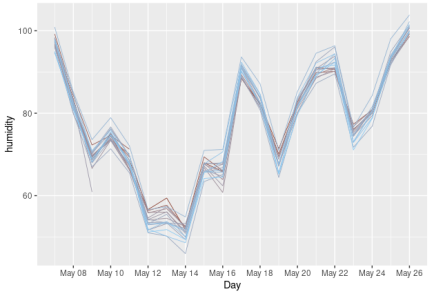
It is shown that humidity and temperature has a strong correlation and it seems to be a linear relationship. The other correlations are less clear-cut.It seems that the maximum temperature first decreases as the height increases to around 35 meters, then it increases as the height increases. We assumed that this is because in the lower height the temperature is more dominant by the ground heat, then as the height increases the sunlight begins to dominate the temperature. We also found that there is a blank in the lower humidity levels when the height is around 35 meters, corresponds to the number in the last paragraph. The reason might be that the lower the temperature it is, the less vapor in the air is evaporated.As the height increases, the maximum value of reflected PAR(hamabot) increases. The lower level of hamabot corresponds to a more variant level of humidity and temperature.

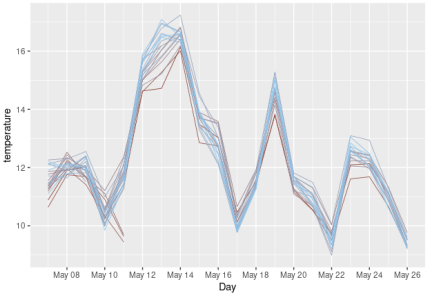
3.2 Predictors associated with Incident PAR

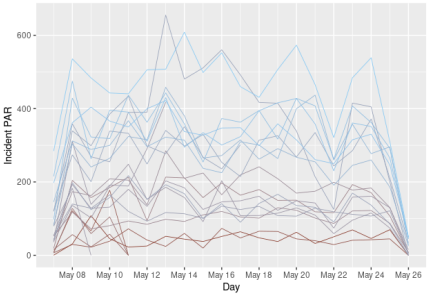
From the scatter plots we can see that there is not direct linear relationship,like the relationship between humidity and temperature, between Incident PAR (hamatop) and other variables. However, we do find some kind of relationships. As the level of hamatop increases, the humidity shifts to lower levels and the temperature shifts to higher levels, while the range of these 2 variables remains generally constant.

3.3 Temporal Trend

Here we use daily means for the value.The range of humidity across all days is around 50% to slightly over 100%, and the humidity generally has a good continuity throughout the days. Combining the height factor, the temporal trend of humidity also has a good continuity and consistency across all height levels.

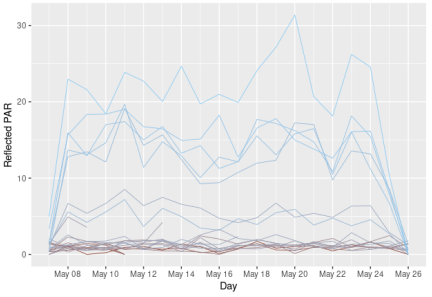






The range of temperature across all days is around 11 degrees Celsius to 17 degrees Celsius, and the humidity generally has a good continuity throughout the days. Combining the height factor, the temporal trend of temperature also has a good continuity and consistency across all height levels.

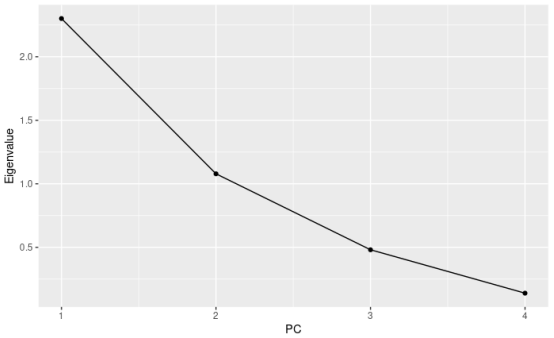
The situation of Incident PAR and Reflected PAR is less ideal. The range of Incident PAR is around 0 to 630 and the range of



Reflected PAR is around 0 to 31. These values fluctuate more drastically and behave differently across different height levels. We think it is kind of strange that the PAR readings fluctuate more drastically.

3.4 PCA

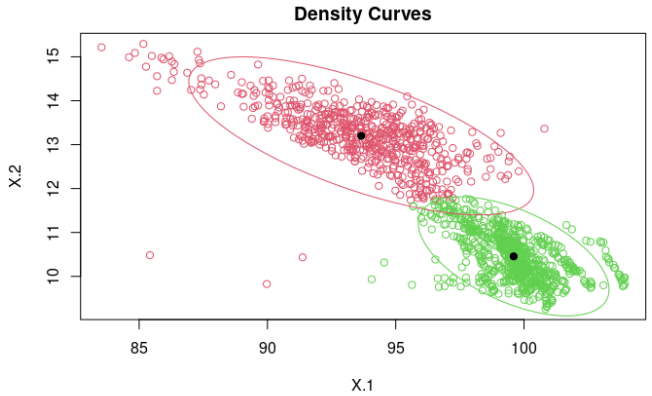
From the scree plot, we know that this data can be approximated by some low-dimensional representation. According to the elbow rule of the scree plot and Kaiser rule, roughly speaking we should take the first 2 PCs. The first 2 PCs contain most information, and are a good low-dimensional representation of the original data.

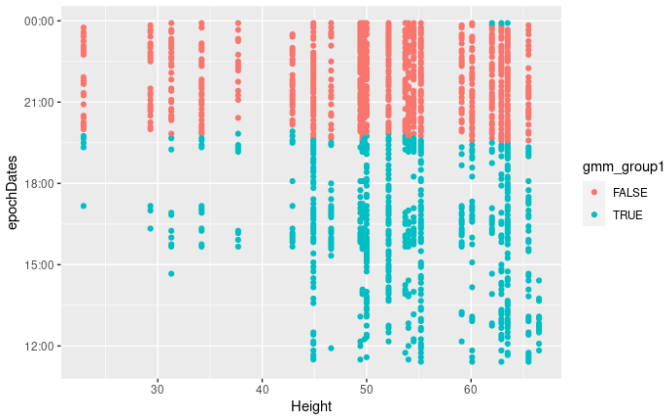


4 Interesting Findings

4.1 Finding 1: GMM

We apply Gaussian mixed model to two environmental variables including humidity and temperature, to distinguish between all observations. From the analysis above, we can see that humidity and temperature trends within a one day period are very similar throughout May, so we choose May 07 as our observations.

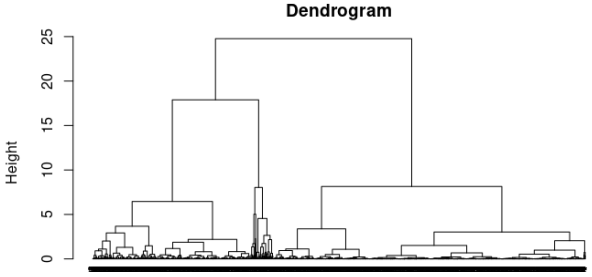




From the Density Curves Figure, we can see that the observations are divided into two groups, one with high temperature and low humidity, and another one with low temperature and high humidity. We then draw a Height-Time scatter plot to examine why observations are divided into these two groups. From the height-epochdates plot, we can see that there is no evident height difference between these two groups. As for the time, observations after 20:00 are mainly divided into group 2 and that before 20:00 are mainly divided into group 1. We figured out that 20:00 is the general sunset time for May. It means that there are some evident characteristics for humidity and temperature before and after sunset: high temperature and low humidity during daytime, and low temperature and high humidity after sunset. It's an interesting and logical finding.

4.2 Finding 2: Hierarchical Clustering

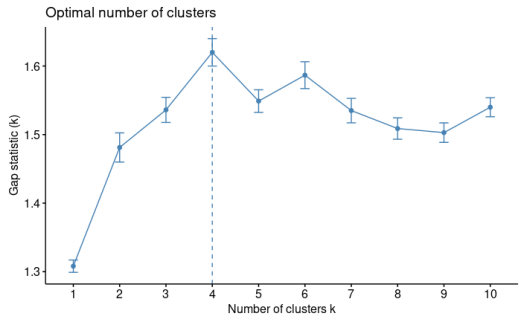
The GMM above clusters according to humidity and temperature, ignoring incident PAR and reflected PAR. However, after applying PCA, we can use the top two principal components (explain 84.6% information) loadings to the cluster, which can present most of the information provided by the previous four variables. We



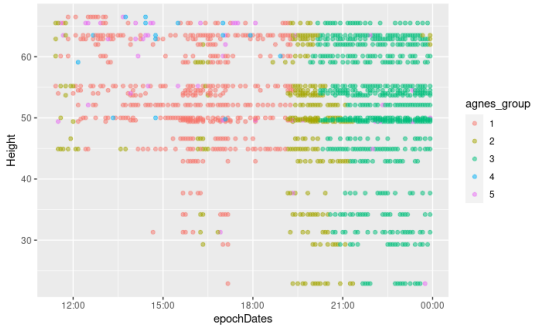
still use data on May 07 as our observations and we use agglomerative hierarchical clustering to divide our observations.

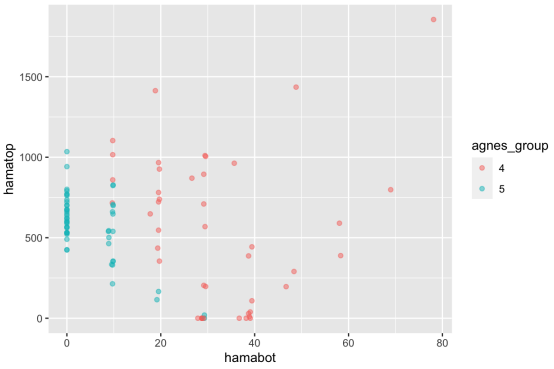
First, we chose a method to determine how close together two clusters are. Among mean linkage clustering, single linkage clustering, complete linkage clustering and Ward's minimum variance method, we can see that the ward's method has the highest agglomerative coefficient 0.9998806. So we train the agglomerative hierarchical clustering model with ward's method. A rough presentation of the tree is shown as the Dendrogram.

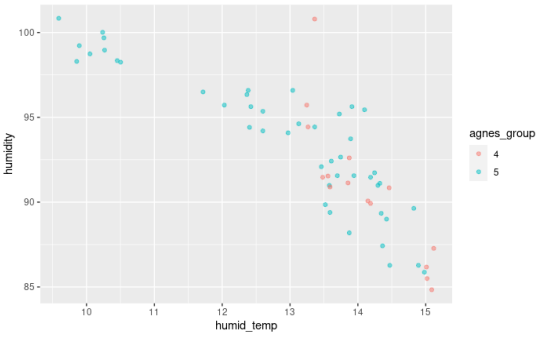
Then we use gap statistic, which compares total intra-cluster variation with their expected values for a distribution with no clustering, to choose the best number of clusters. From the gap statistic plot, we can see that k=5 has the highest gap statistic and is the best, so we cut the tree into 5 clusters and label each observation with the corresponding group.



We draw some plots to see different groups' characteristics of location and time. As shown in figure of epochDates-Height, groups are still mainly divided by time. The hierarchical clustering puts the observation after sunset mainly into group 3, observations between 19:00 and sunset mainly into group 2, and observations before sunset mainly into the remaining three groups. This phenomenon indicates that there's still some characteristics corresponding to the four variables with the various time periods. From the analysis above, we believe that the time stamp 19:00 refers to humidity and temperature and sunset refers to PAR. Besides, among group 1, 4, 5, we can see that the height of group 4 and 5 are all above 50 meters, indicating that height is also a factor to distinguish the four environmental variables of all observations. After examining variables in group 4 and 5, we see from the hamabot-hamatop plot that inflected PAR of group 5 are all below 20 PPFD while group 4 doesn't. Also, group 4 has some observations with low temperature (about 10 degrees) and high relative humidity (about 100%) that group 5 doesn't. These observation's incident PAR are all among 20 and 40 PPFD, and reflected PAR are all 0.

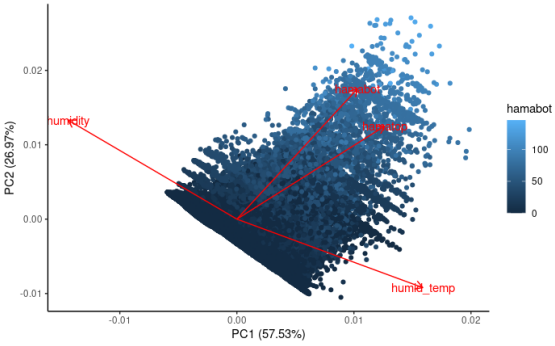






4.3 Finding 3: PCA

We apply PCA to the four variables with all of the observations after scaling. From the summary of PCA in 3.(d), the first two principal components have explained 84.6% information. From the principal component scores, we can see that the first two PCs have similar scores for all of the four variables. Humidity and temperature is a little more important for PC1, and the incident and reflected PAR is a little more important for PC2. As shown in the figure of the PC1-PC2 scatter

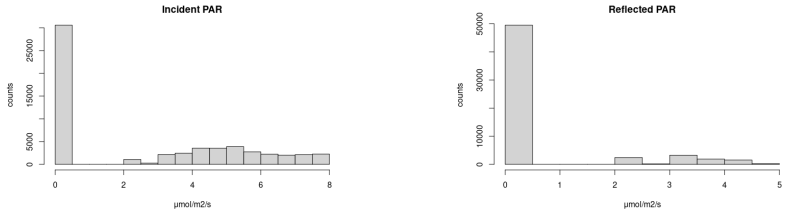


plot, we can see that humidity and temperature almost have the opposite effect on PC1 and PC2. It coincides with the fact that these two variables have a negative linear relationship due to previous analysis.

5 Graph Critique

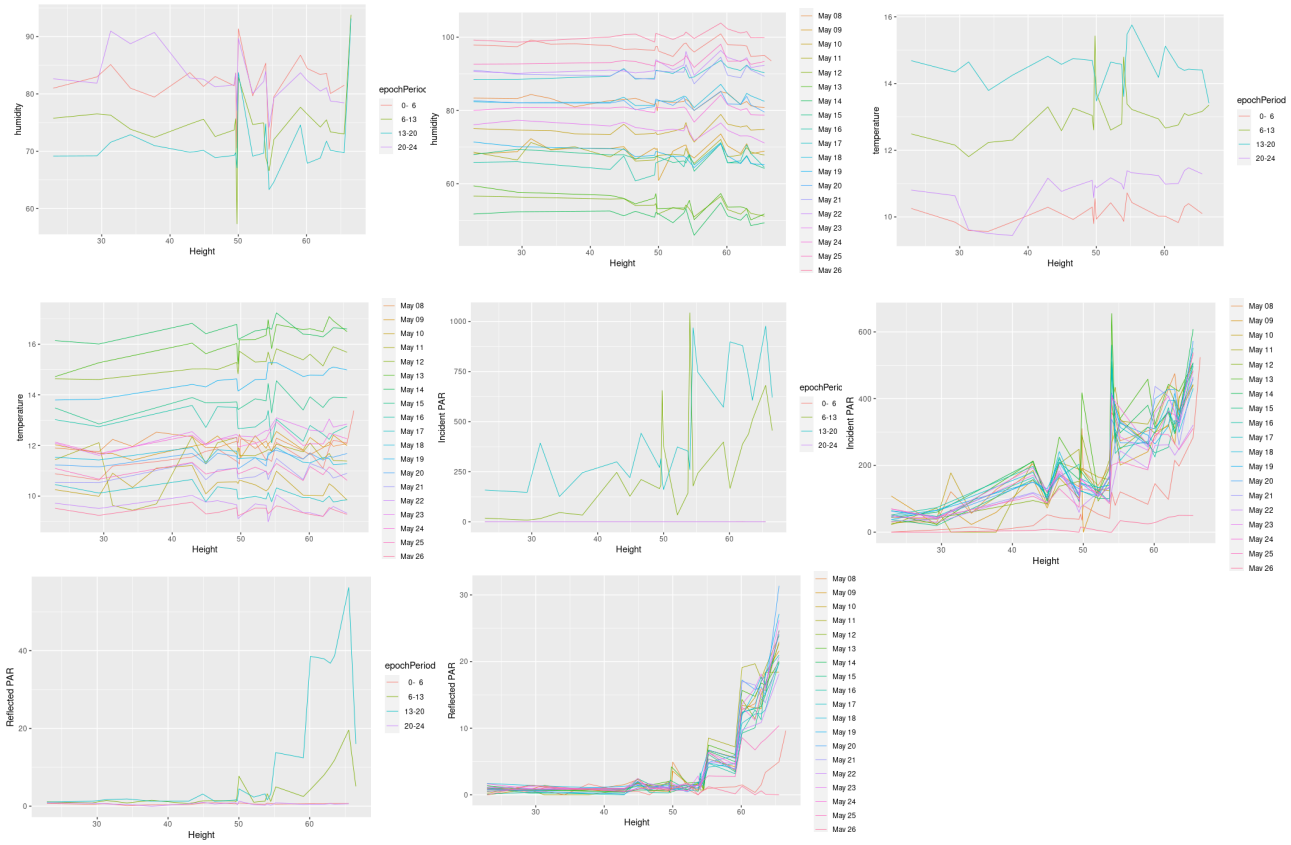
5.1 Log transform of PAR

Since there are many zeros or values close to zero, here we first added 1 to all the data then took the log.



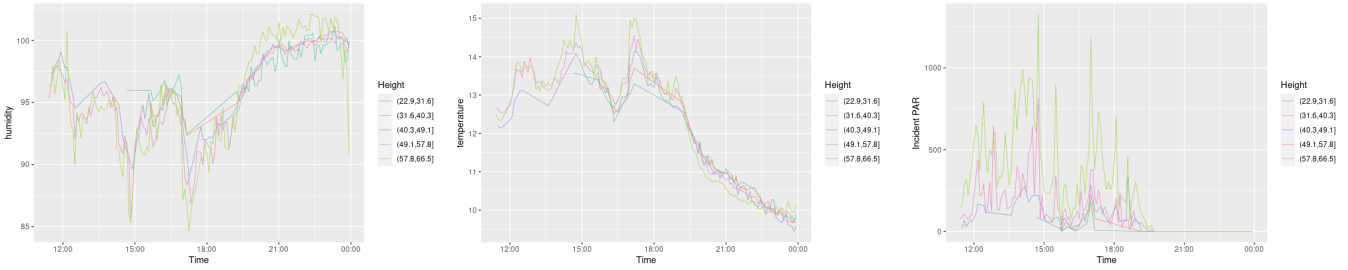
5.2 Figure 3[c]3[d]

We think the boxplots in Figure 3c and 3d are trying to show the distribution of different variables over height. Just as the original paper said, these results suggest that spatial gradients may be present over the height of the tree, but we cannot confirm this without correlating the readings in time. Inspired by Figure4, we set height to be the x-axis and the 4 variables to be the y-axis, then group by certain time period (like before dawn, forenoon, afternoon and night) or days, plot the mean value of different groups together to incorporate the time information. Now we are able to see the spatial gradients more clearly. The level of humidity and temperature varies across time periods, but the trends across heights are similar. Different days also have similar trends. The trends of PAR over height are close and similar across different days but different across time periods.

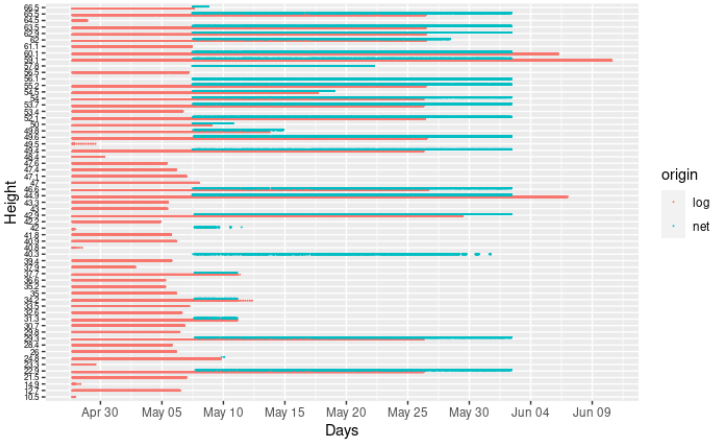


5.3 Figure 4

Here we chose May 7th as the day for observation. The shortcoming for these plots is that there are too many lines in the plot which makes it hard to tell from different lines. To simplify the plot, we divided the nodes into different height groups and then drew the humidity change over time. Then we did the same for temperature and PARs.



5.4 Figure 7



The first thing that came to us is that the fourth plots had many bars and are drawn separately which made it hard to compare. We could draw the two fourth plots together and combine bars with the same height value together, then the plot clearly shows the overlap of two data sets across different heights.

The third scatter plots are even harder to compare. It would be better if they could use a bar instead of a point and draw two plots together grouping by height. The problem with the second plots is that there are too many boxplots which makes it messy. The improvement could be done by dividing the days into different groups and then draw the boxplots.The first plots are the worst, we did not get what they are trying to convey.