PROJ1 write-up

Xinyi Sheng, Yucan Zeng

2022-10-11

Abstract

While the original 2005 paper was a preliminary study of the macroclimate of Sequoia trees along the California coast using wireless sensor networks, this report thoroughly examines the Sequoia dataset differently, focusing mainly on the preparation and cleaning before starting the data analysis. There are five sections: raw data collection, data cleaning, data exploration, interesting findings, and conclusions.

1. Data Collection

a.Paper Summary

The thesis focuses on the ecophysiology of Sonoma redwood trees in California using wireless sensor networks to explore the microclimate's complex spatial variability and temporal dynamics around coastal redwood trees. Air temperature, relative humidity, and photosynthetically active solar radiation (PAR) were monitored every 5 min for 2 m for 44 days during the early summer of the redwood tree's life by placing magnetic suction sensors on a 70 m high redwood tree. The paper describes the placement of magnetic suction sensors, one sensor every 2 m from 15 m to 70 m above ground. Since the canopy is thicker on the western side of the redwood tree, the sensors are located on the western side. Sensors were also placed at 0.1-1.0 m from the trunk to ensure that the trees' microclimate trends could be captured directly.

Moreover, TASK and TinySQL software serve as the framework for collecting and processing the data. The first phase investigates the extent and distribution of the data independent of time and space by transforming a 3-dimensional dataset with a temporal dimension, a spatial dimension, and a dimension related to the sensor values themselves into a 1-dimensional dataset; the second phase adds a temporal dimension to investigate the temporal trend of the data can be seen in the weather over the time scale with a temporal gradient. The third phase investigates the spatial trend of the data based on the 1-dimensional dataset. The third stage examines the spatial trend of the data based on a 1-dimensional dataset, where the trend in height helps to derive a theoretical model of canopy density and the existence of a spatial gradient in tree height. The last stage examines the three-dimensional data, and we contrast the sensor values by shades of color in different times and spaces. Through humidity graphs, the dense canopy can have a buffering effect on the lower part of the tree.

b.

i.Sensor Deployment Mode

Since early summer contains the most dynamic microclimate variability, we decided that sampling all sensors every 5 minutes would be sufficient to capture this variability for a month. tolle placed the sensors from 15 m above ground to 70 m above ground, with a spacing between nodes of approximately 2 m. This spatial density ensured that we could capture the gradients in sufficient detail to interpolate accurately. Simultaneous placement of the envolop starts at 15 m because most of the leaves are in the upper part of the tree. We primarily measured

the western side of the tree, which has a thicker canopy and provides the greatest buffer against the direct effects of the environment. Not only that, but we also placed the sensors 0.1-1.0 m away from the trunk of the tree. These nodes were placed very close to the trunk to ensure that we captured the microclimate trends that directly affect the tree rather than the broader climate. figure 1 in the Tolle article shows the final position of each node in the tree. We also placed several nodes outside of the angular and radial envelopes to monitor microclimate near other previously installed biosensing devices.

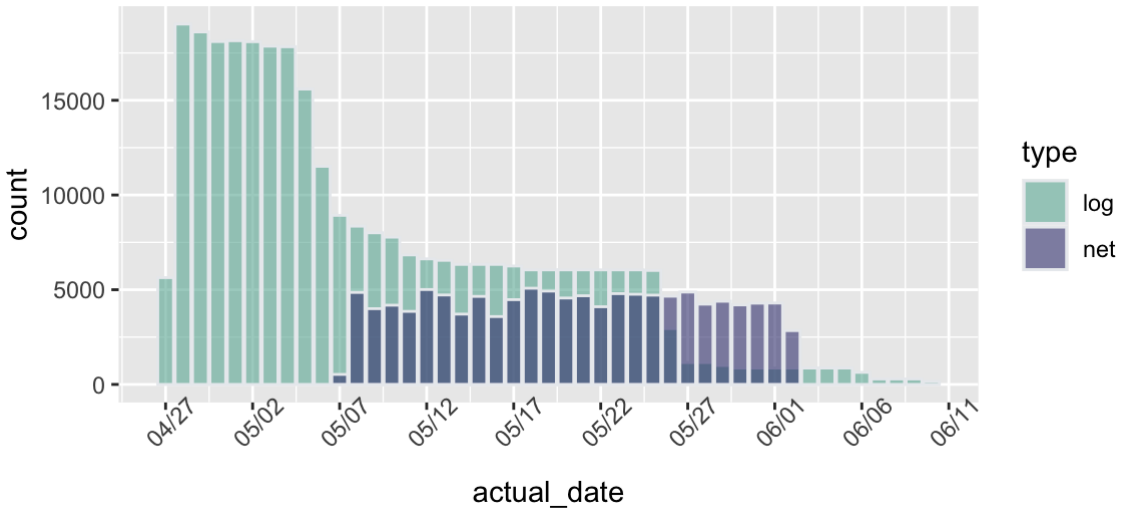
ii.Summary of the Dataset

The collection includes measurements made over 44 days, from April 28, 2004, to June 10, 2004, in Sonoma County on two trees (the interior and the periphery). Since the edge tree only has around one-fourth as many data points as the interior tree and the edge tree was not covered in the original publication, this report will mainly concentrate on the inner tree with 33 sensor nodes. Every 5 minutes, the 33 sensor motes, which were mounted on the redwood tree at a height of 15 to 70 meters, took readings for humidity, temperature, PAR (photosynthetically active radiation), and reflected PAR. Each mode's log files on the disk included the collected measurements, which were also sent to researchers through the GPRS network (net data). Prior to deployment, each mote was calibrated to guarantee the precision of the measurements. Mote-location-data.txt is a text file that stores each sensor mote's profile. Each mote's height, angular position, and distance from the stem were noted, allowing for a multi-dimensional study and subletting to the interior of the tree.

2.Data Cleaning

a.Check Histogram

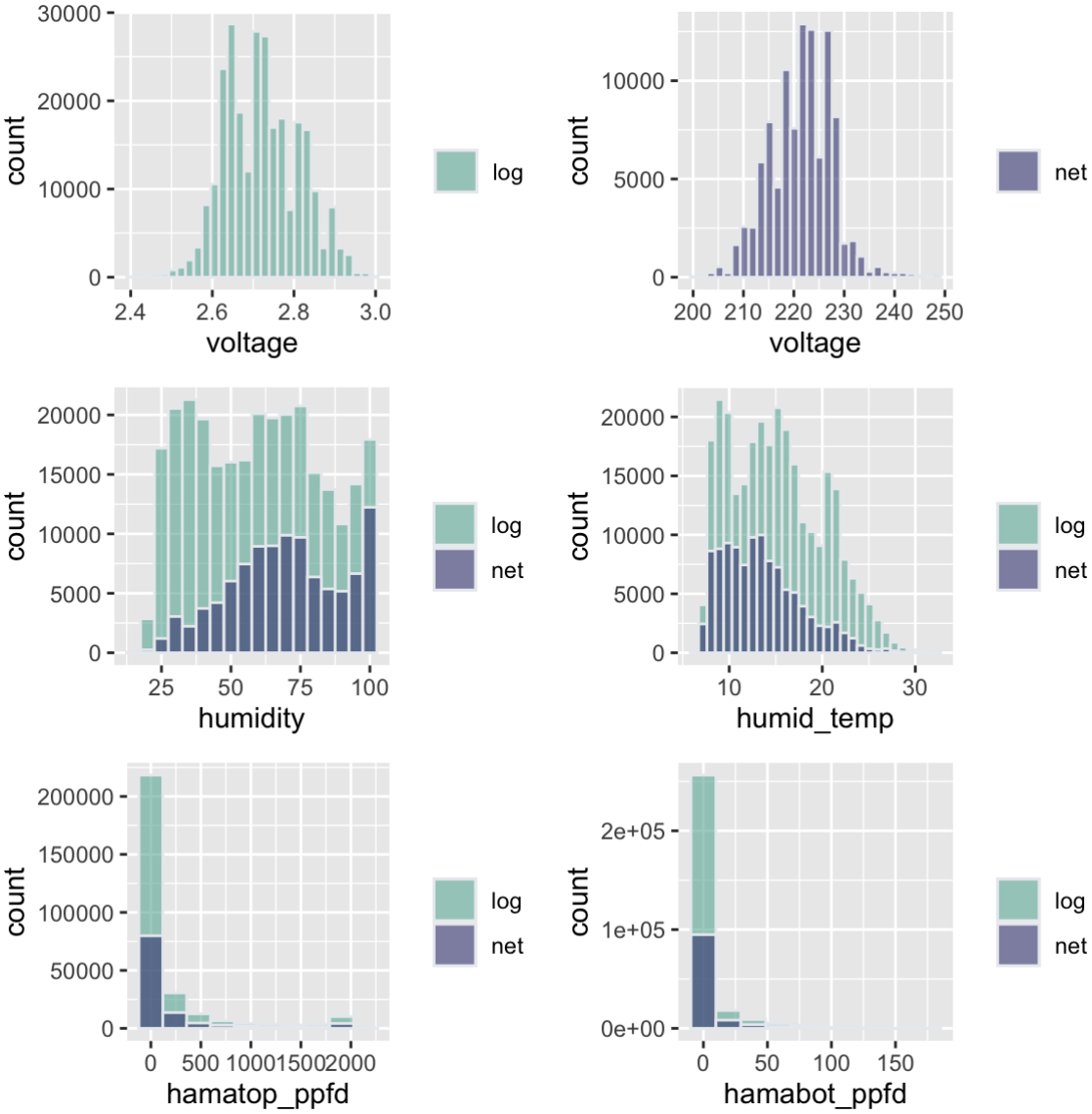
Before plotting the histogram, we initially filtered the variables according to the range in Tolle's table1 for each sensor and range of voltage, removed the NA, controlled the local log voltage at (2.4, 3), network voltage at (200, 250), then controlled the temperature at (6.6, 32.6) and humidity at (16.4, 100.2). Also, because the time is a trash, we also corrected the time.



Data Distribution over Time by Log and Net

We use the raw data to plot the number of measurements recorded by loggers and networks for each period. Figure 1 shows that the network data are missing and only appear from May 7 to June 2. Locallog data decreased on May 5 and May 6, and on May 26. Next, we made histograms of the number of measurements recorded by the

logger and the network for each period for voltage, humidity, temperature, hamatop, and hamabot, respectively. It was obvious that in the voltage distribution, the network data was very problematic, and the locallog data was concentrated around 0. And we found that the value 200 to 300 was taken when filtering the network because if we go beyond this range we can only see the locallog and the network value was not useful. Also, we can clearly see from the locallog and network two graphs that the units of voltage are not the same, so we performed the operation of converting the voltage. In the humidity histogram, we can see that locallog records data evenly distributed at each humidity value, but network logger recorded more data at humidity greater than 50, and we speculated that network logger may not be sensitive enough for smaller humidity. In the temperature histogram, both locallog and network data were well recorded. Since hamatop and hamabot only recorded data when there is light, it was reasonable that most of the data of hamatop and hamabot were recorded near 0.



Count of Data

b.Remove Missing Value

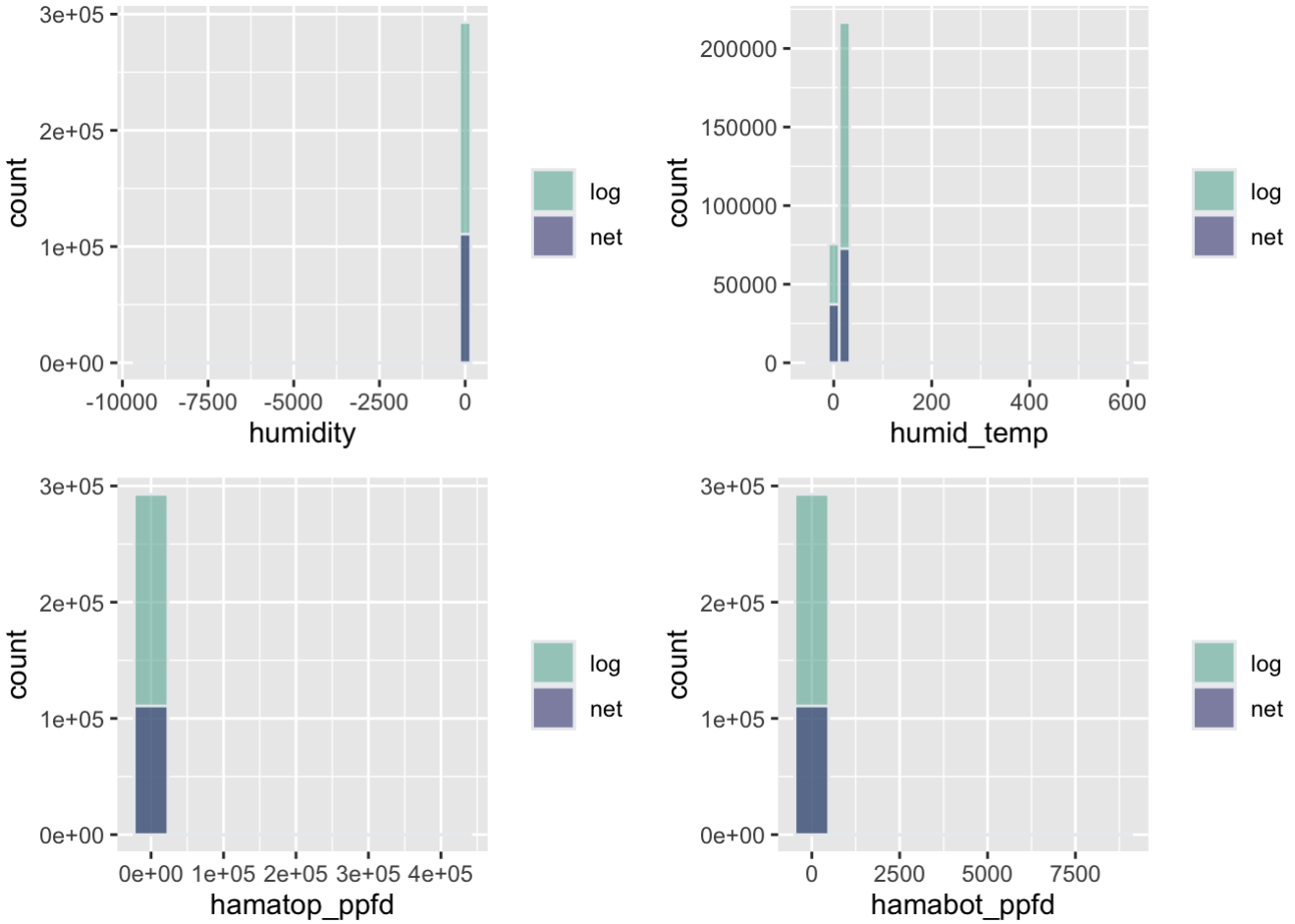
After viewing the dataset we preprocessed data by removing NA, outliers, duplicates and filtering the ranges of the variables. First, we dropped the rows containing NA, and then we filtered the data for duplicated rows of epoch and nodeid, keeping only one row. Since the network voltages were converted to ADC, we divided the

network voltage data by the conversion factor. Next, we replicated the voltage filtering steps outlined in Tolle's paper, removing entries with voltages greater than 3V or less than 2.4V, and filtering the data according to the data range in table1 in the paper.

c.Incorporate the Location Data

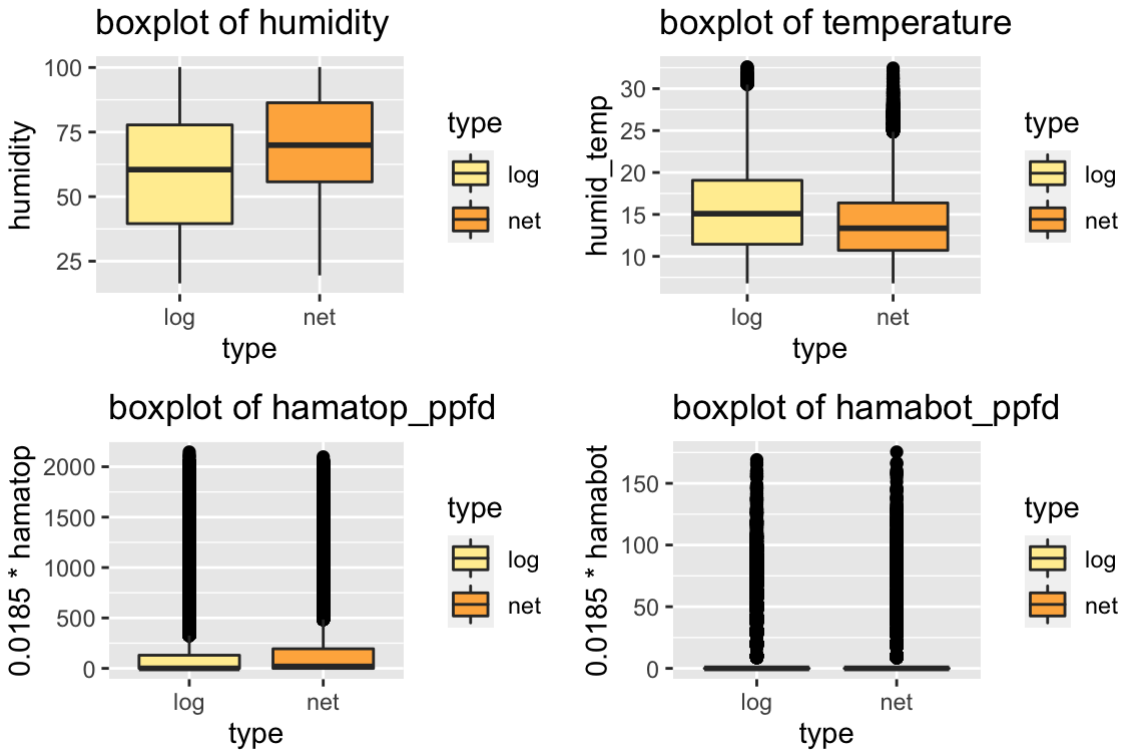
After removing the outliers, we filtered out the anomalous sensors using "moto-location-data". We used inner join by 'nodeid=ID', and then we got a dataset with 18 variables. Our last dataset has 14 variable.

d.Identify Outliers



Histograms

Tolle's article mentions the problems associated with outliers corresponding to the measurement range of humidity, and we also found when looking at the histogram data that humidity less than 0 and temperature up to 100 degrees are obvious outliers that do not correspond to reality, so we dropped these data.



Boxplots

According to the boxplots, for humidity, we filtered only the readings with quantile at (0, 0.02). Secondly, for temperature, we filtered out the quantile at (0.98, 1). And for hamatop and hamabot, we also filtered out the quantile at (0.98, 1). In the humidity boxplot, the mean value of network is larger than the mean value of local log exactly in line with our guess that the network sensor is not sensitive to lower humidity and may miss a large number of readings. In the boxplots of hamatop and hamabot it is obvious that the readings are heavily concentrated around 0, so the boxplots would have more outliers. Because there is a large amount of time without sunlight, resulting in 0 readings, we do not remove them considering the specificity of the data.

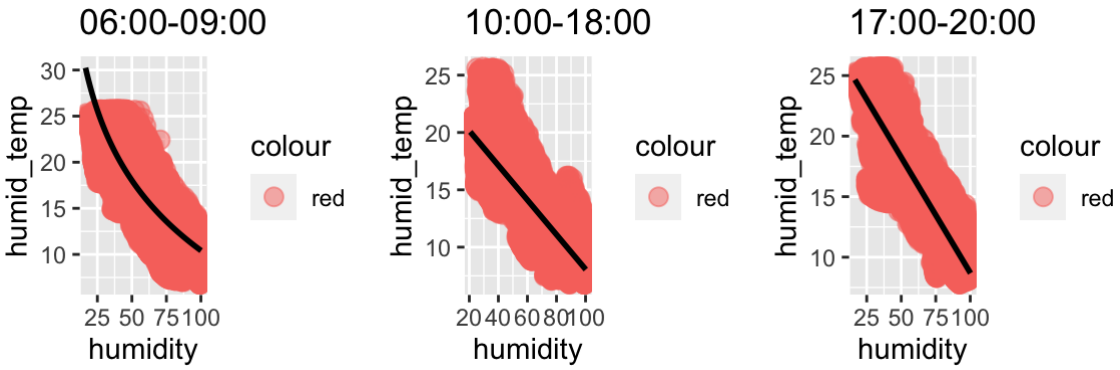
e.

In addition to using the filtering range given in the article, we also converted the battery voltage of the network sensor by a coefficient during data processing, thus further removing the outliers instead of simply dropping that part of the data or leaving it unprocessed.

3.Data Exploration

a.Pairwise Analysis

We choose three different time periods: 6 am to 9 am (sunrise time), 10 pm to 18 pm, and 17 pm to 20 pm (sunset time). Because it is clear in Figure 4 of the Tolle's article that there are spatial and temporal trends in humidity and temperature during these three time periods, and intuitively, these two variables change dramatically at sunrise and sunset, and we wanted to find and quantify the relationship between temperature and humidity. At the same time, humidity and temperature show dramatic fluctuations at noon, so we explored the potential correlation between humidity and temperature during the three time periods.

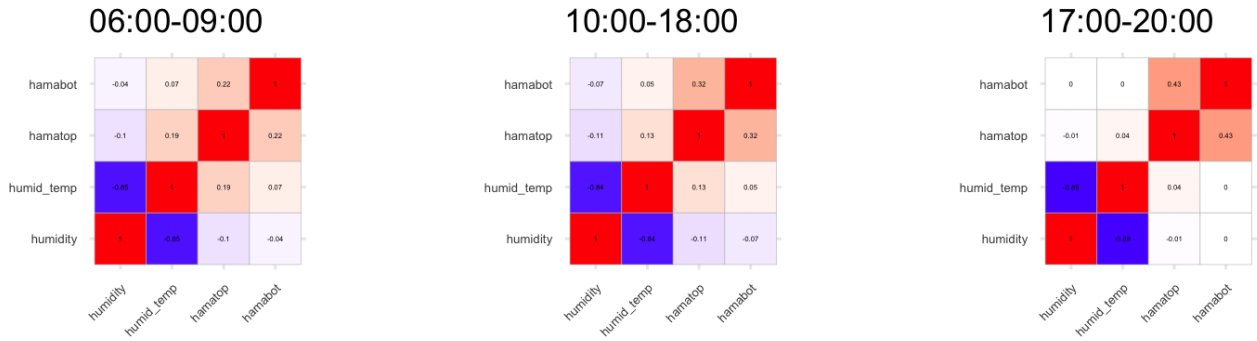


Scatterplots

The pairwise scatter plot of humidity and temperature shows a relationship between temperature and humidity. As can be seen, when the temperature rises, the humidity decreases. There is a linear relationship between temperature and humidity at sunrise and noon. However, the trend is logarithmic at sunset.

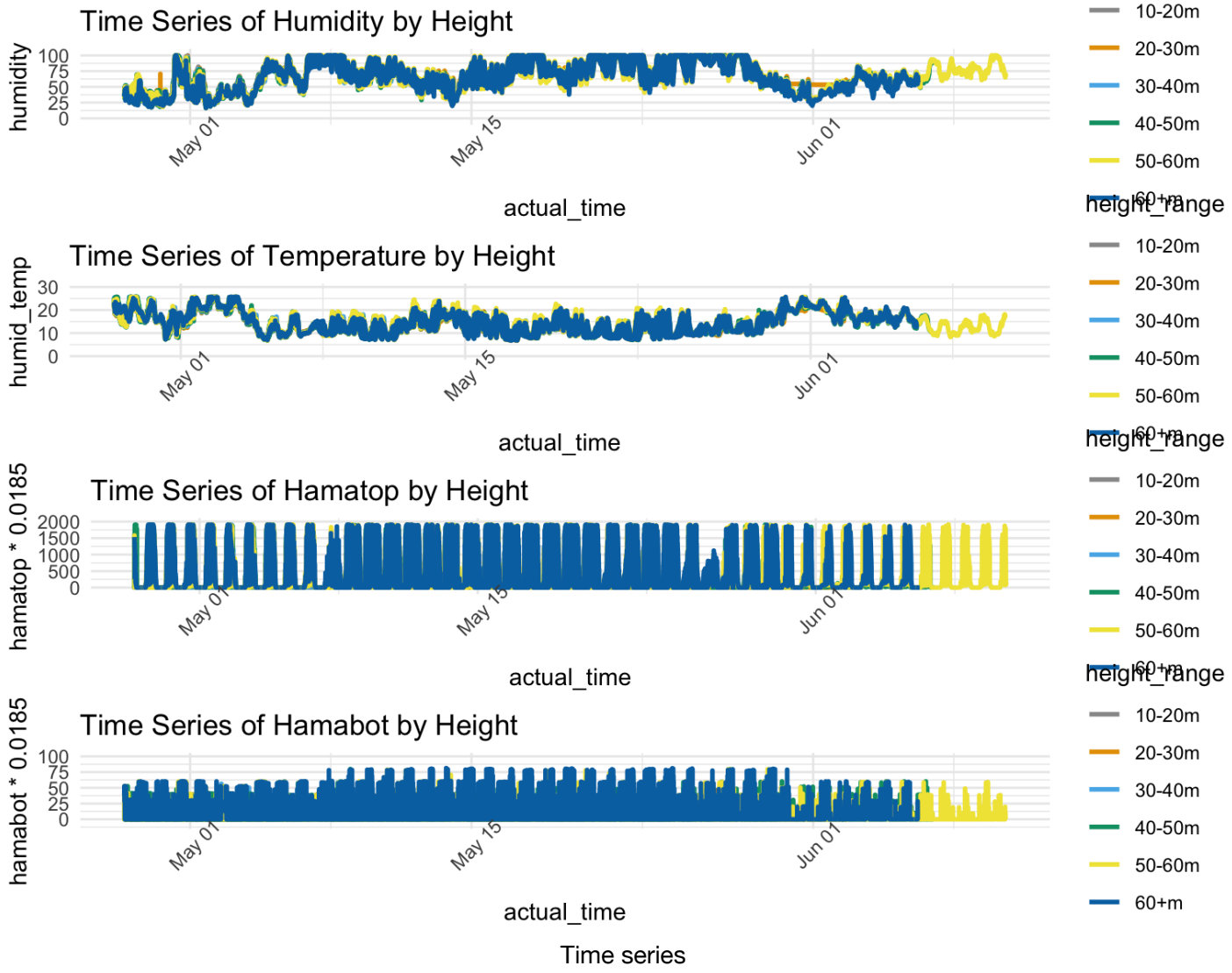
b.Correlation

We created correlation plots based on the three time periods mention above to recognize any predictors associated with incident PAR (hamatop). It is evident that Incident PAR has a medium correlation with reflected PAR (hamabot), and the closer they are to sunset, the more robust the correlation is. In addition, Incident PAR has a weak correlation with temperature, and the correlation coefficient is more significant during the day than at night.



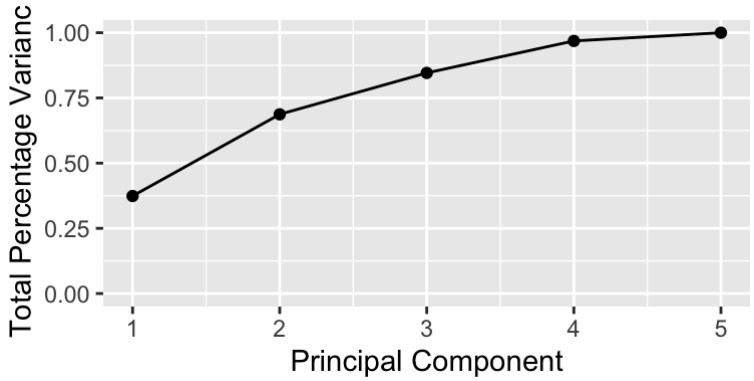
Correlation plots

c.Time Series

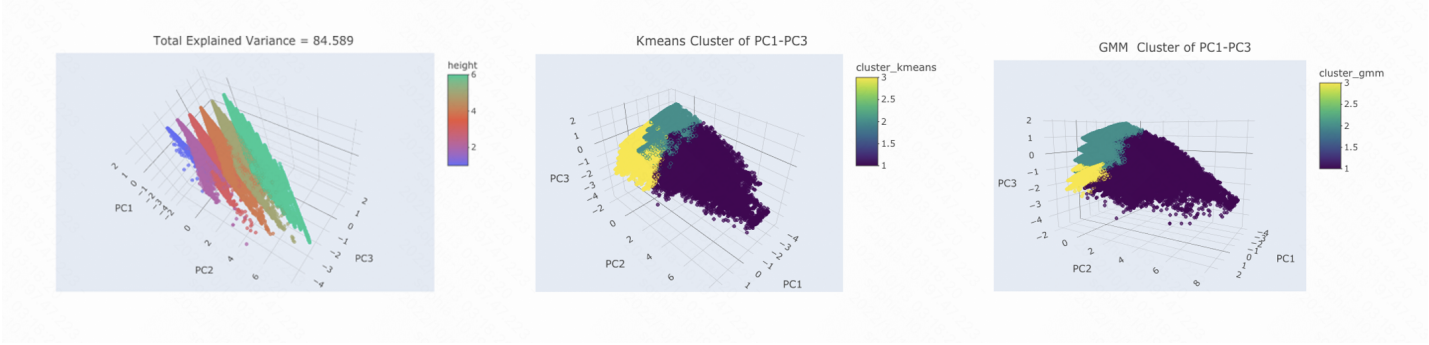


We want to evaluate the daily cycles of climatic measures after looking at the hourly pattern of measurements within a day to determine if those readings were noticeably changed from April 27 to June 10. The temperature was not steady during the investigation period, as can be shown from time series plots. The average temperatures in May's first and last weeks were much warmer than those in the middle of the month. Because of the large climatic change over time, it may be difficult to generalize the findings, and the data from May 1st in the earlier portion of the investigation may not be accurate. Additionally, the structure of the humidity time series was roughly inverse to the temperature. We also can still see that both the incident PAR and the reflected PAR had a clear seasonal pattern that also corresponded to how sunlight behaves. The incident PAR and reflected PAR, behaved consistently from day to day with the exception of May 7 and May 27, and when PAR and RPAR were much lower, possibly due to rain or clouds.

d.PCA



Scree Plot



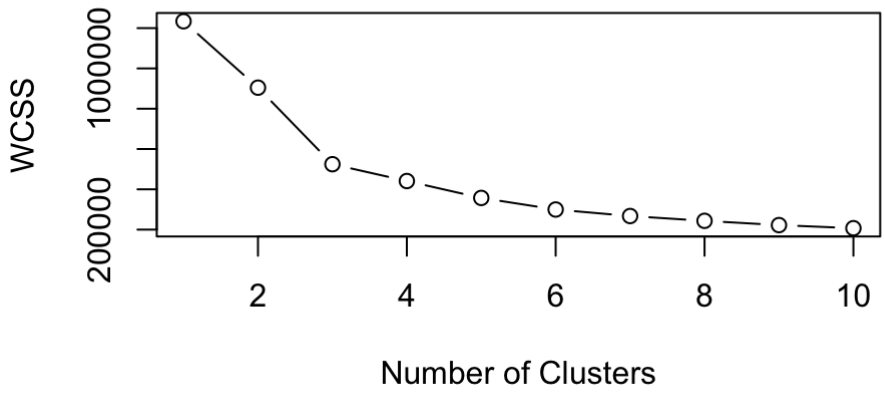
PCA and Clustering

We used PCA analysis for dimensionality reduction on these five variables, and from the scree plot, we could see that PC1 to PC3 had explained more than 80% of the variance, and indicated that there might be lower dimensional representation, so we considered using PC1 to PC3 for cluster analysis in the next study. We found that temperature and humidity were more dominant in the PC1 direction, while height, hamabot, and hamatop were more dominant in the PC2 direction. Also we found that the more sensors there are, the more samples are recorded at higher heights, mainly concentrated in the 40-50 m height range. So we guess that the top of the tree presents a harsher environment and is more vulnerable to weather than the bottom which is protected by denser foliage.

4.Interesting Findings

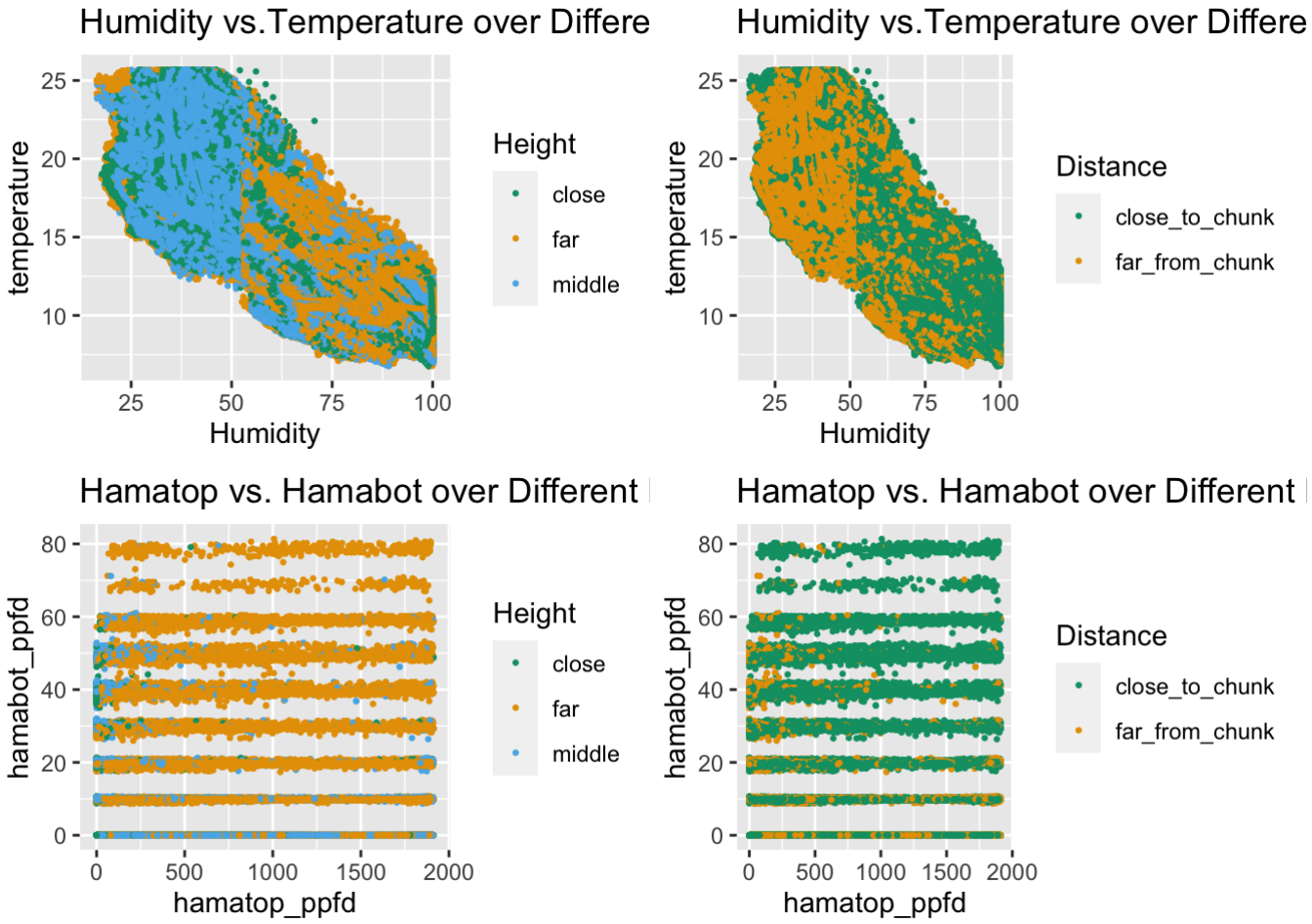
finding 1

We reduced five features into three denoted by PC1, PC2, and PC3 by cluster, and divided them into six layers according to height range, which shows that PCA is appropriate for dimension reduction in this case. We found the elbow by the WCSS method and decided that it should be divided into 3 clusters. To verify the clustering performance, we use k-means and GMM. As can be seen in the figure ‘PCA and Clustering’, the results obtained by these two clustering methods are not consistent, but it is obvious that they are divided into 3 clusters, which indicates that our PC1 PC2, and PC3 have very strong interpretability.



Clusters of Clients

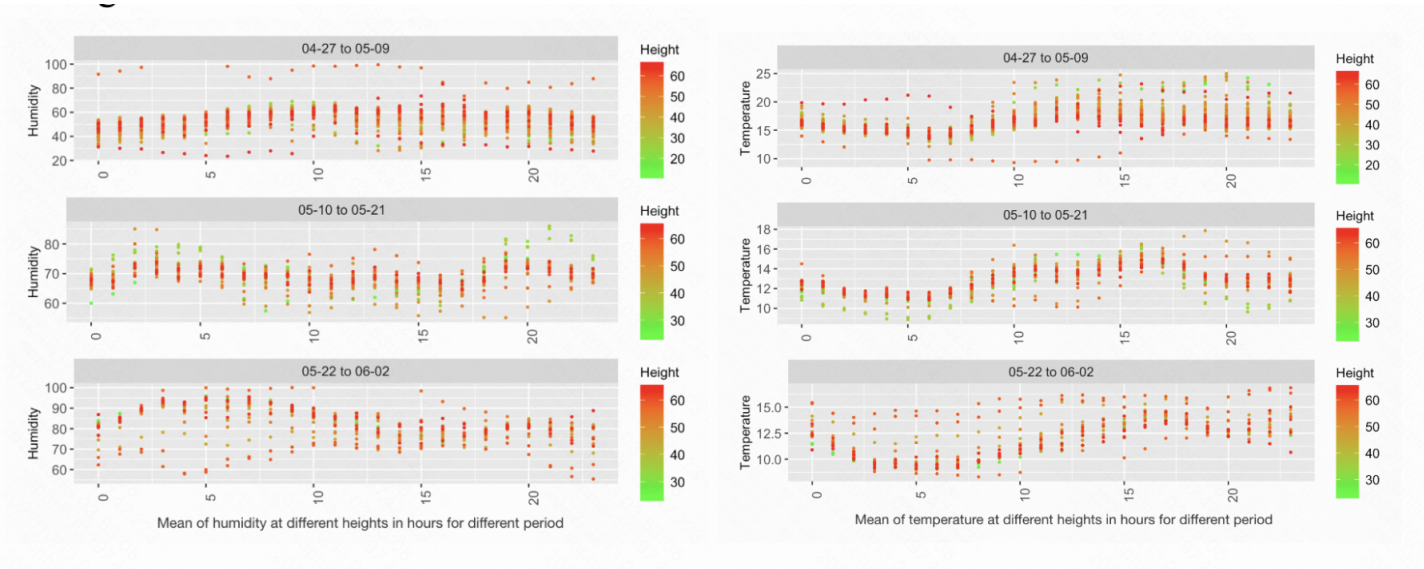
finding2



At different heights, humidity and temperature are negatively correlated. For sensors placed higher, lower humidity will usually as the change of height and distance higher temperature. In our data exploration, we hypothesized that temperature and humidity are negatively correlated. We set above 50 m as high, 40-50 m as middle, and below 40 m as height close to the ground, and we placed sensors less than 1 m from the trunk as close and greater than 1 m as far. Figure A shows the relationship between temperature and humidity and height. When the temperature is low, these points are dominated by yellow, and most of the sensors are placed in high places,

which also indicates that the humidity is high. When the temperature is high, the dots are predominantly blue and the sensors are mostly placed at low locations, which also indicates low humidity. Figure B shows the relationship between temperature and humidity and sensor placement distance. When the temperature is low, these dots are predominantly green and the sensors are mostly placed close to the tree trunk, which also indicates higher humidity. When the temperature is higher, the dots are mainly yellow and the sensors are mostly placed far from the tree trunk, which also indicates higher humidity. In addition, we also found that when the sensors were placed high, more sunlight was incident on the tree, indicating that the leaves at higher locations received higher sunlight and therefore grew more densely, while more sunlight was reflected near the tree trunk.

finding3



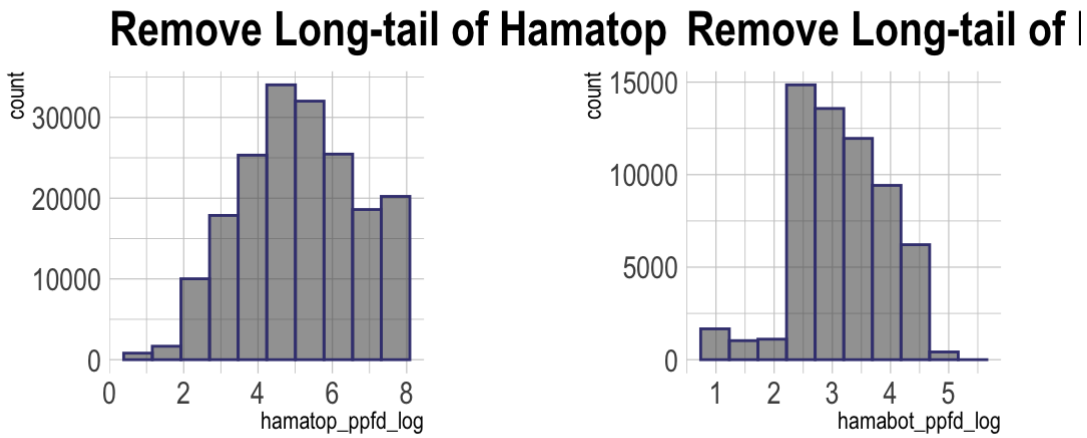
The graph shows that, particularly, the temperature was high and the humidity was low around May 3 and May 30. Since temperature and humidity vary considerably between different period, we divided 37 days into 3 time period: 4/27 - 5/9, 5/10-5/21, and 5/22-6/2, and calculated the average temperature and humidity every hour during these three time period to obtain 6 different hourly trends. Although the average humidity and temperature levels were different for these 3 periods, the change patterns were similar. Basically, the temperature started to increase and the humidity started to decrease from around 7:00 am, the temperature reached its maximum and the humidity reached its minimum around 3:00 pm, and then the temperature gradually decreased and the humidity gradually increased around the sunset time. However, from April 27 to May 9, the daytime humidity is higher than the evening humidity, and then from May 21 to June 2 the daytime humidity is lower than the evening humidity. When we divided the data points by height, it did not appear to be a spatial confounder of temperature and humidity, except for the period April 27 through May 9. During this period, height was positively correlated with temperature and negatively correlated with humidity. And there were fewer sensor nodes working between May 21 and June 2, so the average points for that period were not as stable as before.

5.Graph Critique in the paper

a.

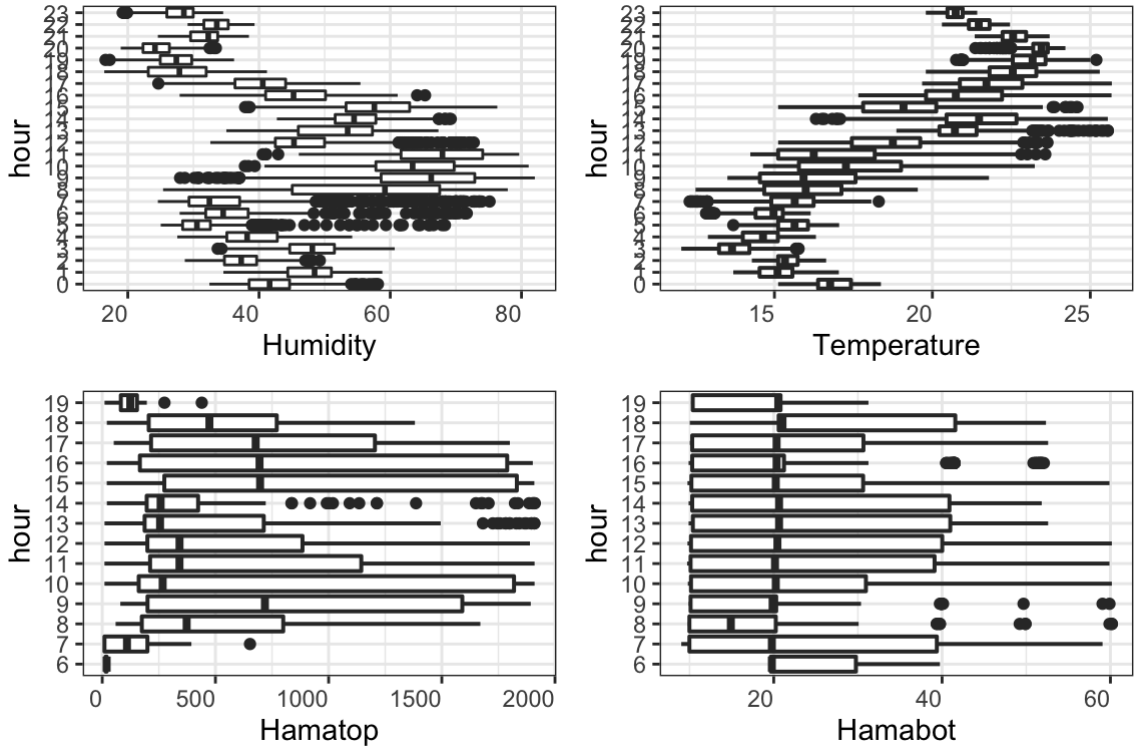
Figure 3 [a] in Tolle's article examines the univariate distributions of temperature, humidity, incident PAR, and reflected PAR. Although these graphs clearly show the distributions, the temperature plots are too widely spaced on the x-axis, resulting in a unimodal distribution for temperature. At the same time, there may be multiple modal

distributions around 10, so we removed the long-tail of incident PAR and reflected PAR by using a log() transformation.



b.

3 [c] represents the horizontal box plot distribution of the measured values over time. The graphs can be seen to be unconvincing due to temporal mixing effects, but the problem is resolved in 3d. However, the authors plot height as a categorical variable at equidistant distances on the y-axis, so we think it should be clarified to the reader to prevent the misconception that y-values are proportional to height. 3[d] eliminates the time effect by differencing the measurements at each node from their mean and plotting them in a box plot with height. Although these plots actually provide more information than 3[c], the title and axis text of 3[d] are too small to read and see because there are still many gaps in the horizontal direction.



Distribution of humidity, temperature, hamatop and hamatop in differen hours

The boxplot in 3(c) shows the relationship between heights and the value of temperature, relative humidity, incident PAR and reflected PAR respectively; the boxplot in 3(d) shows the relationship between heights and the difference from the mean of temperature, relative humidity, incident PAR and reflected PAR respectively. All plots in two figures show that at the high position of the tree, the incident PAR and reflected PAR tend to have a high value, and thus a higher temperature. Correspondingly, the higher the temperature, the faster the water vapor, resulting in lower humidity. However, the plots in Figure 3 do not convey the complete message. In Figure 4, we can see that these values also change as the time passes. Thus, we made boxplots and showed the relationship between time and the value of temperature, relative humidity, incident PAR and reflected PAR. Plots show that at day time, specifically 8am. to 18pm., there is relatively higher incident PAR and reflected PAR, and thus should have a higher temperature but lower relative humidity. Thus, one can easily assume that the temperature at the bottom of the red wood should be relatively low, and will increase as the height increases, and relative humidity viceversa. However, this assumption only holds on heights ranging from 20-40 m. At the height ranging from 40m to 60m, the temperature does not increase significantly and the relative humidity keeps a high value, which violates our naive assumption. Thus, there should be an internal regulation mechanism in red wood that is worth further analyzing.

c.

The left plot in Figure 4 shows the temperature, humidity, incident PAR, and reflect PAR over time on May 1, 2004. Each line indicates a different sensor with a different color in the first two plots. In addition, there is a vertical line marking the sharp change in slope. These plot should add a legend that indicates different colors to represent different nodes, because if there is no legend, then the plot should use only one color. In the plots of PAR and reflected PAR, the report does not assign different colors according to the different sensors, which is inconsistent with the first two plots. On the right of Figure 4, the scatter plot shows the relationship between the four measurements and the height variable. The scatter has two different colors, which should be specified in the legend in addition to the scatter plot. In general, the plot tries to convey the relationship between measurements, time and height. We should also reduce the range of the y-axis, using ylim to make the range smaller.

d.

It is possible to combine network and local log data and plot them on the same graph using two different colors.

Conclusion

In this report, we first briefly describe the data collected by Tolle et al. (2005) and extensively clean the dataset. Next, we explored possible findings during the dataset search. Finally, we explained the cluster in detail by k-means and GMM, we study the incidence and reflection of light at different heights, widths, and interior tree and edge tree, and we study the hourly trends of temperature and humidity at three different time periods to come up with interesting findings.