Redwood Climate Analysis

Jincen Li, Yi Xu

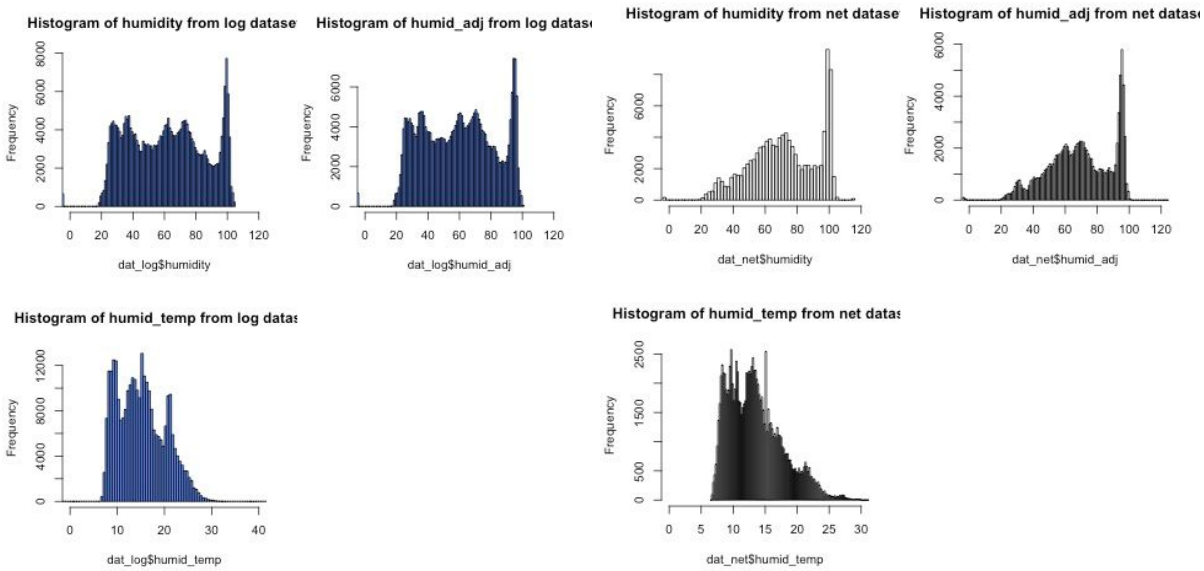
1. Data Collection

(a) This paper reports on a case study of microclimatic monitoring of a coastal redwood canopy where is located in Sonoma California. The researchers studied the data point collected by a wireless sensor network called "macroscopes" that recorded 44 days in the life of a 70-meter tall redwood tree, at a density of every 5 minutes in time and every 2 meters in space. The variables collected by each node on the net include air temperature, relative humidity, and photosynthetically active solar radiation. Each data point collected by this sensor network can be viewed as having a location in three-dimensional space: a time dimension, a height dimension, and a dimension corresponding to the sensor value itself. By conducting a multidimensional analysis methodology, the researchers were able to reveal trends and gradients in this large and previously-unobtainable dataset. Through the continuous data collecting and analysis, the result of the study showed how dynamic the microclimate surrounding a coastal redwood is and verified the existence of spatial gradients in the microclimate around a redwood tree. Furthermore, by deeply understanding the dense and wide-ranging spatiotemporal data obtained from the macroscope which can be possibly applied on a large-scale processes of carbon and water exchange within a forest ecosystem; at the same time, having captured enough data to track the changes in these gradients over time, they can begin using this data to validate biological theories which was hard to achieve by the ordinary technology.

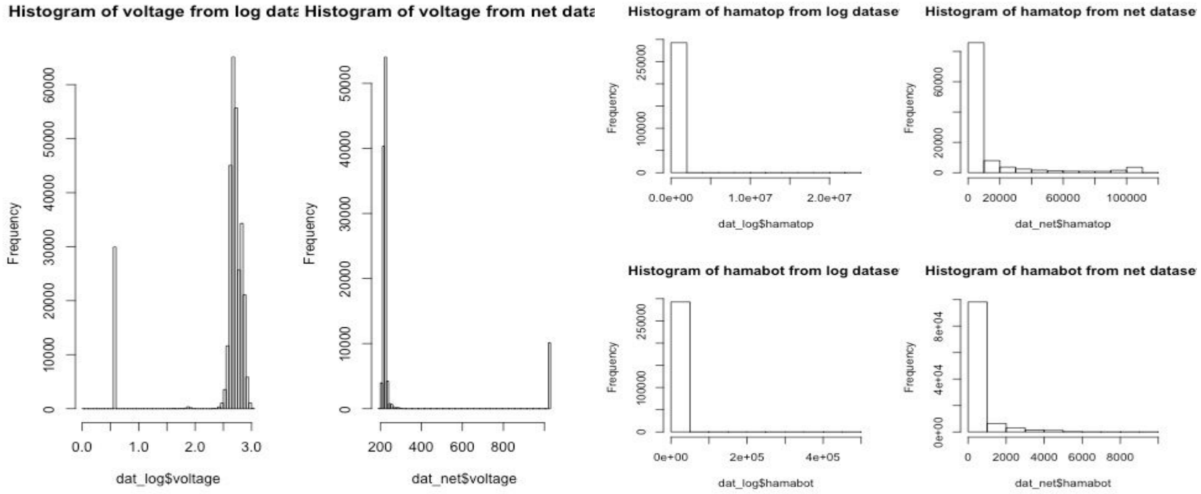
(b) Gathering data on the environmental dynamics around 70-meter tall redwood tree for 44 days requires robust system design and a careful deployment methodology. The researchers selected a suite of sensors and designed packages with software application that allows that sensors access to the environment while resisting the elements. One month during the early summer, sampling all sensors once every 5 minutes. The duration of the data recording in the early summer they chose was every 5 minutes which would be sufficient to capture the variation in the most dynamic microclimatic season. The nodes were placed very close to the trunk to ensure that they were capturing the microclimatic trends that affected the tree directly, and not the broader climate, and the vertical distance was 15m from ground level to 70m from ground level, with roughly a 2-meter spacing between nodes. This spatial density ensured that they could capture gradients in enough detail to interpolate accurately. They also placed several nodes outside of their angular and radial envelope in order to monitor the microclimate in the immediate vicinity of other biological sensing equipment that had previously been installed. Collecting high-quality real-world data with a wireless sensor network requires a comprehensive deployment strategy that carefully tests and calibrates the sensors prior to deployment. The main variable they measured is the traditional climate variables – temperature, humidity, and light levels. Temperature and relative humidity feed directly into transpiration models for redwood forests. Photosynthetically active radiation (PAR, wavelength from 350 to 700 nm) provides information about energy available for photosynthesis and tells about drivers for the carbon balance in the forest. The standardized temperature and humidity sensing performed in a shaded area with adequate airflow, implying that the enclosure must provide such a space while absorbing little radiated heat while the sensors measuring ambient levels of PAR must be shaded but need a relatively wide field of view. And for the deployment methodology, the strategy includes two calibration phases: roof and chamber. The roof calibration provided a real-world data source for the PAR sensors: direct sunlight. The purpose of the chamber calibration phase was to understand the response of the temperature and humidity sensors to a wide range of phenomena.

2. Data Cleaning

(a)



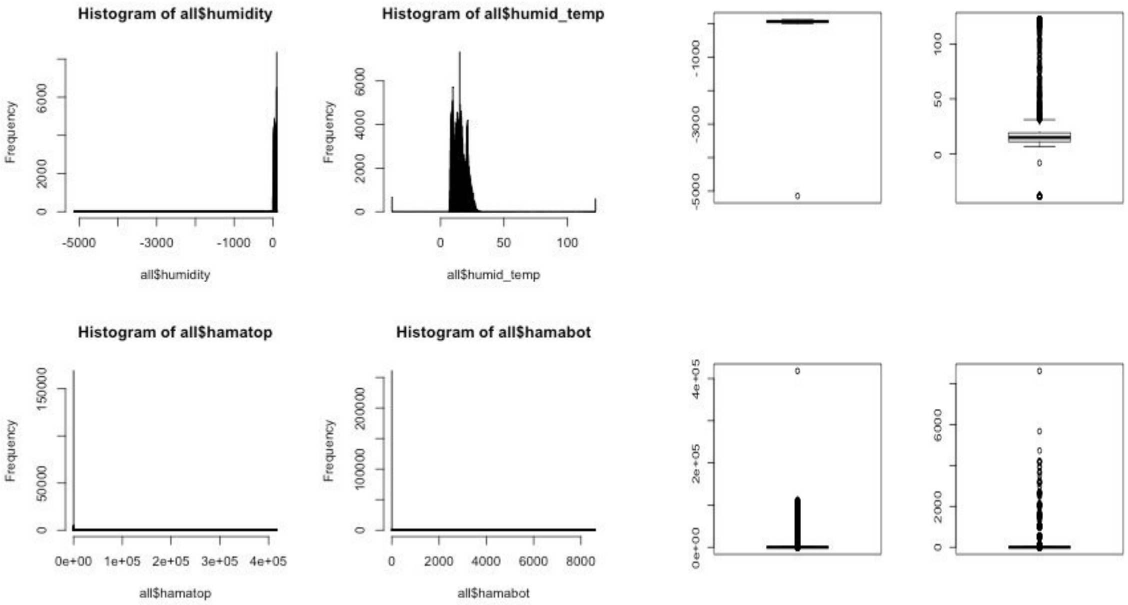
There are a total of 11 variables in both net and log dataset including measurable variables and the time collected from the research forest. We select "humidity", "humid_temp", "humid_adj" for histogram comparison based on our interests since the temperature can be an essential factor to observe in the redwood forest and the humidity can be directly affected by it. By comparing the same variable in each graph, we see the distribution for each variable from the net and log data is different (especially for “humidity” and “humid_adj”), but the range for each variable in two graphs is basically consistent.



Then we further check other variables for consistency which the voltage, hamabot and hamatop stand out. In log dataset, the range of voltage start from 0 to 3 while the voltage range of net lie between 200 to 1000 which indicates the inconsistency of the unit of the variable. By checking the Analog to Digital Conversion, we determine the conversion factor and multiplying it on voltage from net dataset, we have the variable consistent within the same range. Similarly, we notice that hamabot and hamatop are not consistent in two data files, so by checking the unit conversion table, we multiple 1/54 on the hamatop from log dataset since the direct light source is sunlight and then we have consistent hamatop. The same applied to hamabot.

First we omit all the missing value from the main table (all dataset). We found the total number of missing measurement is 12532, and the variables the data missed are “humidity”, “humid_temp”, “humid_adj”, “hamatop” and “hamabot”. Furthermore, the date of missing data is continuous which starts from 2004-05-07 19:19:58.713061 to 2004-05-13 13:20:32.009966 which indicates the data recorded from the sensor is missing the observations (we have not corrected the wrong date for the log data).

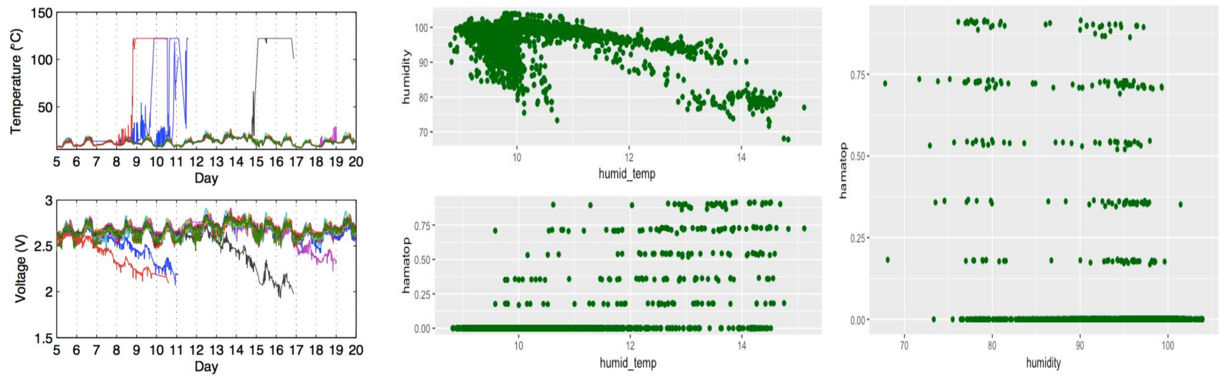
We merged the location data with the main table by matching nodeid. Now we have 15 variables in a new dataset including “Height”, “Direc”, “Dist”, and “Tree”. The new dataset provides more comprehensive information about the design of the system. We further clean the data by removing duplicate information.



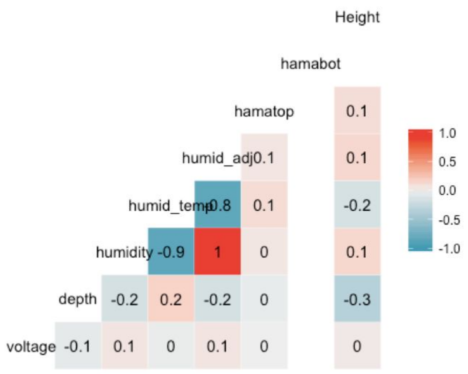
Here we have the histogram of the four variables comparing with their boxplots. We can see nasty outliers which lead the plots showing extreme range. Usually, we define the data point outside the 3*sd range as outliers. Here we use the boxplot for identifying outliers. We constructed a function to identify the data value outside 1.5 times the interquartile range above the upper quartile and below the lower quartile. The data value identified by the function will be removed as outliers.

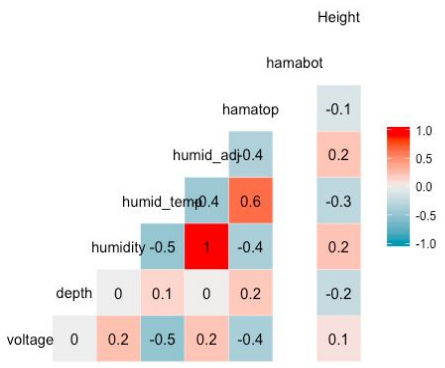
We remove an outlier if it's significantly outside of the range, and if we have a lot of data like the size of this project, we can also remove questionable outliers since the sample response is not critically affected by few data points and increase accuracy of measurements.

3. Data Exploration

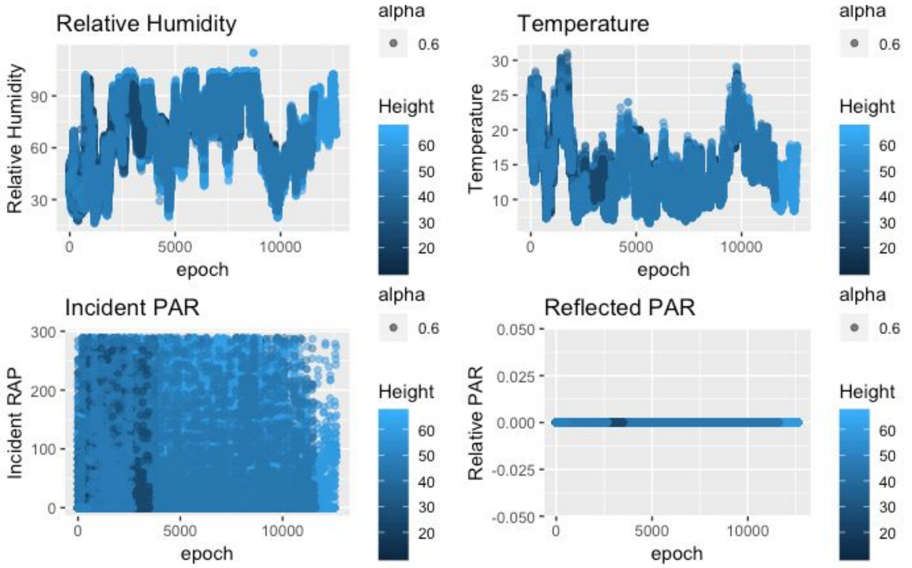


The time period we choose is 2014/05/05 to 2014/05/08. From figure 6 on the paper, we see the temperature outliers are correlated with battery failure. Therefore, we choose a period with relevant stable voltage which should have less outsider factors that affect readings on other variables. Compare the temperature with humidity and hamatop both plots shows some correlations. There is a rough decreasing trend showing on the scatter plot of temperature and humidity, the humidity is relevant higher with lower temperature. By checking with hamatop and temperature, we see the readings of hamatop are basically assigned into 5 categories. There shows an increasing trend on categorical level when the temperature increased. Since both variables show correlations with temperature, we further select hamatop with humidity to see any associations within this two and the scatter plot does not give us much correlation information, the spread of point are evenly distributed with some questionable outliers.

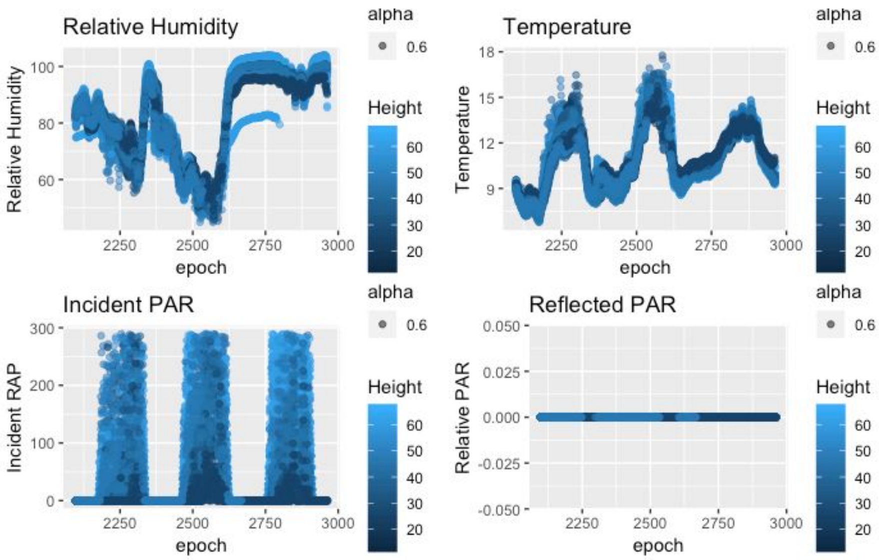




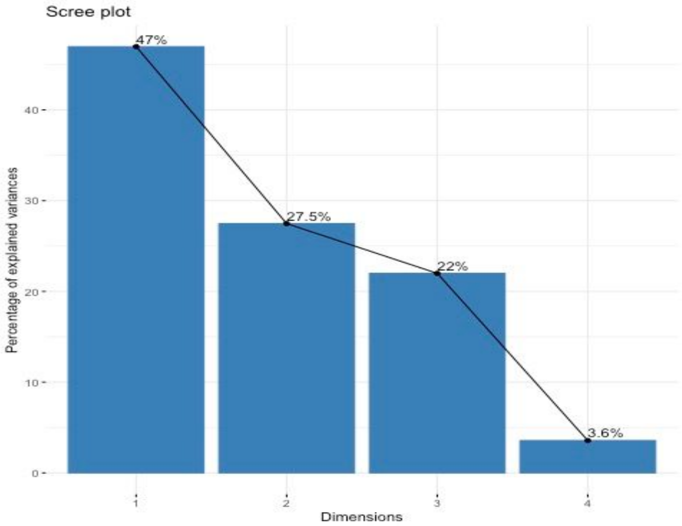
From the correlation plot, we see the clear correlations between each variables. Here we want to check if there is any predictors associated with Incident PAR which represented by hamapot. Graph at left gives the information of the cleaned main table, and the graph on the right shows the correlations on the day we selected. Basically incident PAR does not associated with any predictors in general, and it has higher positive correlation with humid_temp on the day we selected which may due to chance error.



We consider each variable as a time series and look at its temporal trend. The graphs we plot show the time and value of four variable (humidity, humid_temp, hamatop, hamabot). First we draw the time series for the entire experiment. There is a suspicious outlier in the humidity plot, and the overall oscillation is roughly symmetric regardless the height. For the temperature variable, there is seasonality characteristic on small range duration showing in the temperature data. The temperature through the whole experiment is relevant low in the beginning of the experiment which is in April. And for Incident PAR and Reflected PAR, we couldn't see much information from the pattern. The readings of Reflected PAR are all zero, and the reading of lower height nodes centralized at the beginning of the experiment with full range Incident RAP readings.

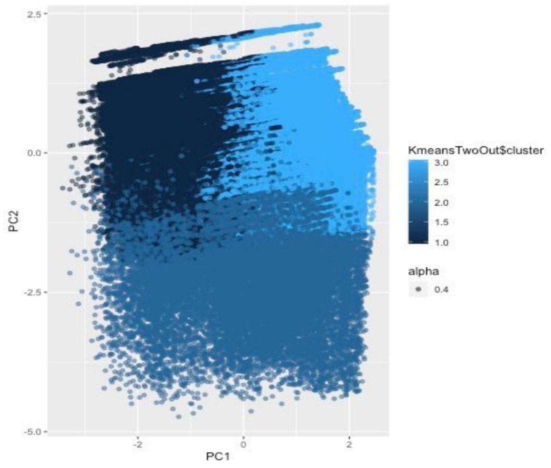


Now we choose a consistent time period with 3a for better comparison. Here we convert the real date and time with epoch expression. The reading is recorded every 5 minutes in a day, and every epoch represents 1 duration of reading. By doing simple math, we have the time range from epoch 2098 to 2962 which gives us the data from 5/5 12 am to 5/9 12 am. The time series of humidity shows a rough increasing trend over time and there is a significant drop down in the middle of the three days with a separate branch of high height. As for temperature, by looking at a smaller range for the experiment, the seasonality is more clearly to see, the fluctuations have some similar patterns. The reading of Incident PAR are clearly not continuous, and the readings of Incident PAR are small with lower height nodes. Finally, there is a strange finding on the Reflected PAR data, the readings are zero which consistent with the overall data plot, but the height of nodes goes down over time which may relate to the spatial gradient.

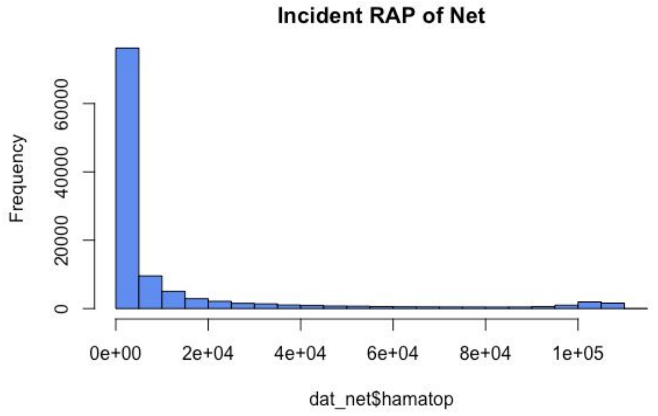


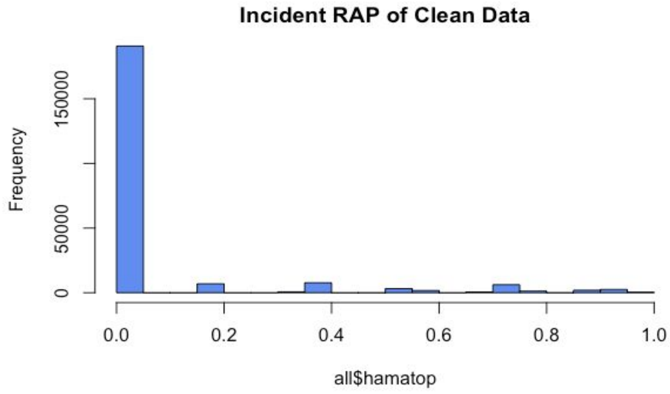
From the PCA analysis, we can see the PC1, PC2 and PC3 included the majority of the variation. It indicated this data can be approximated by some low-dimensional representation.

4. Interesting Finding



When we have three centers for the K means analysis, it gives us the best visualization of cluster blocks that shares the closet characteristics from the 4 variables. We tried center number 2 and 4, which both don't have well-defined clustering as center number 3 (both have unexpected too many overlappings). This number happens to be the representative reduced dimension from the PCA analysis. The intuitive idea behind the matching number we think is that with dimension 3, the PC captured most of the variation from the data and presented sufficient level of data for interpretation. While cluster is a technique for finding similarity groups in a data, there are 3 majority characteristics these 4 variables can be splitted into without too much interaction with each other. We may also see the 3 majority characteristics as the main variations the variables spread into, therefore there is an interesting number match between the reduced dimension of PCA analysis and the best-fit center number of K-means.

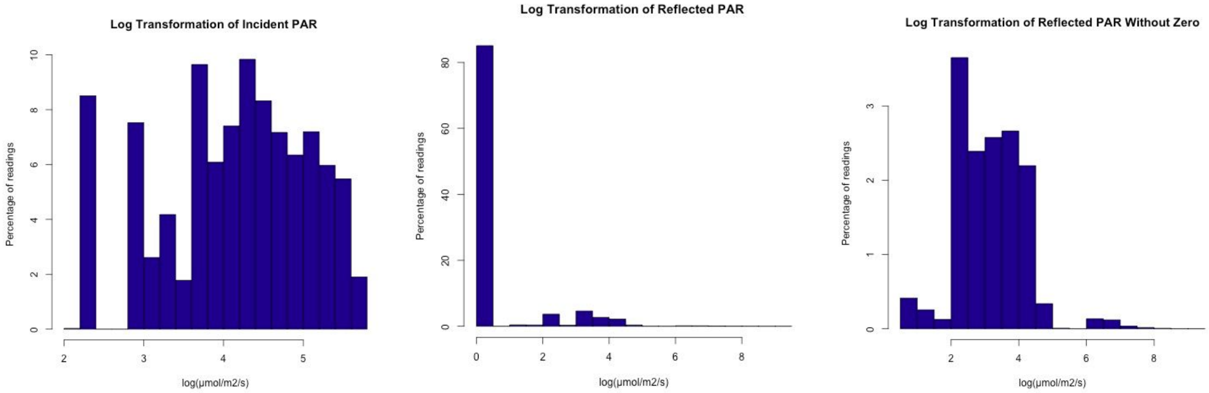




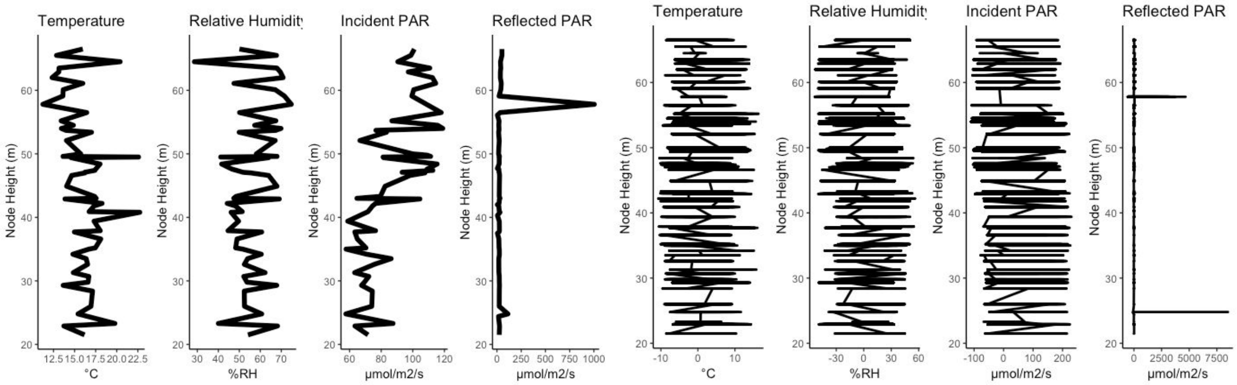
We see from the histogram that the readings of Incident RAP was continuous of the raw data collected from the sensors. After cleaning with the outliers and unit conversions, we see the readings are assigned into several categories. This also showed in the scatter plot and the time series from data exploration that the readings of Incident RAP are divided into groups with cleaned data.

From the time series we plot for the entire experiment, we found nodes with high height have less data readings at the end of the experiment.

5.Graph Critique in the Paper



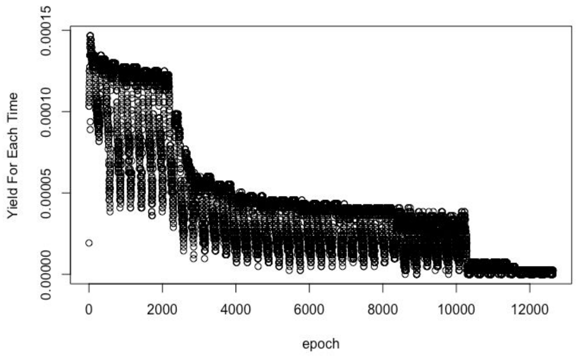
By looking at the log transformation graph of the incident PAR and reflected PAR, the incident PAR graph got improved, however, the reflected PAR still has a long right tail since it contains a large number of zeros. Thus, we deleted the data with zero hamabot and then did the log transformation and plot.

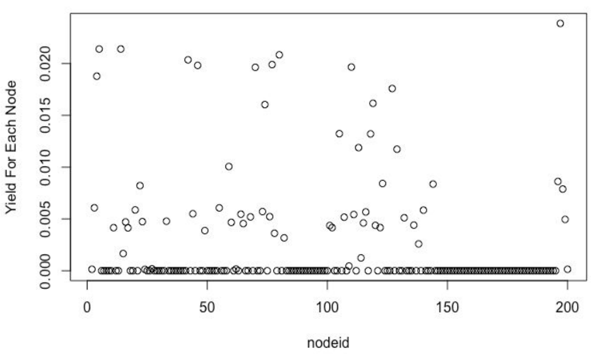


The boxplot in Figure 3[c] and 3[d] try to convey the spatial trends on each variable. Figure 3[c] shows the distribution of all the readings taken by each sensor at each height, and Figure 3[d] removes the temporal variation and focus more closely on the spatial trends. And by examining the distribution, we can determine whether there is an absolute relationship or a tendency. By comparing Figure 3[c] and 3[d] with four height plots on the right side of Figure 4, they both show that temperature and humidity change in a nonlinear way over height, and both incident and reflected PAR charts show the top of tree receives more lights, while the reflected PAR decrease more quickly than the incident PAR does. However, the data source for Figure 3[c] and 3[d] is the entire dataset of this experiment, and Figure 4 focuses on a selected date May 1st. Indeed, selecting a short period of data, like a day, would be a better way to show the spatial trends of each variable, since the weather can change during the experiment, affecting the distribution of the variable readings. One-day data removes the random weather variation and focuses closely on the change of variables over different height. I tried to plot the mean for each variable instead of plotting all the data so that the trend can be presented clearly. Also for the incident and reflected PAR, I tried to delete zeros since the paper says in the PAR readings, we can see a spatial trend in both the mass of the distributions and in the outliers. And by comparing PAR plots, we noticed there is a extra peak for reflected PAR in the readings difference from the mean.

The first two plots in Figure 4 is difficult to visually read especially near the two endpoints of the graphs. Each line represents the reading taken by an individual sensor throughout the day. And the temperature and the spread of temperature throughout the tree move in a similar way, as well as relative humidity and the spread of relative humidity, therefore line stack together and it’s hard to distinguish each line. I cannot distinguish all the colors in these two plots. Moreover, as mentioned in the paper, we cannot see whether the change of graph shape is due to a particular trend over space or random local variations in air movement. In order to create a better plot, we could try to add some transparency to the graph. And instead of plotting the data from all sensors, we can randomly select a few sensors at each height and then make the plot. The plot is easy to read and can still show the trend.

Defined on the paper, the yield is calculated by the way that if each mote at each timestep reports any data, then consider it to report to a 1, otherwise report to a 0. However, by discovering the experimental dataset and analyzing the data’s consistency, we realize that not all data recorded by mote are useful. Like outliers, and those temperature recorded under the extreme voltage are not reliable, and should be cleaned up. Therefore, considering the mote to report to a 1 without determining if the information is useful or not is too coarse. In order to generate a better visualization, we should determine how useful the data is first and then pick 0 or 1 according to the usefulness. Moreover, the x-axis for graphs in Figure 7 is days, and the yield data is recorded at each timestep. However, the paper doesn’t clarify how to determine timestep. Is it a day? Or is it just the 5-minute period? If each time period is defined as a 5-minute interval, and we say there is a mote that works only one time during an entire day, and we still mark a 1 to this mote since it reports some data. Clearly, the yield from this mote is unreliable and should be abandoned. Therefore, we treated clean data as the useful report. Yield for each time point is calculated by summing the total number of each epoch from the clean data and dividing the number of installed motes. Yield for each node is calculated by summing the total number of each nodeid from the clean data and the total number of time points. Below is the graph I got. The left side one is the graph showing the yield percentage at each time point, which indicates that the yield decreases with time increasing. The right side graph shows the yield for each node. The yield for each node is mainly zero. In order to generate a better visualization to highlight the difference between network and log data, we should plot both network and log data on the same graph with some transparency, distinguished by different color and shape. Therefore, the difference can be directly presented on the graph.


Yield For Each Time
0.00000
0.00005
0.00010
0.00015
0
2000
4000
6000
8000
10000
12000
epoch


Yield For Each Node
0.000
0.005
0.010
0.015
0.020
0
50
100
150
200
nodeid