Lab 1 - Redwood Data, Stat 215A, Fall 2018

9/13/2018

Introduction

The paper "Macroscope in the Redwoods" presents a case study of a wireless sensor network 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. Each node measured air temperature, relative humidity, incident photosynthetically active solar radiation (PAR) and reflected PAR. The author wanted to analyze the complex spatial and temporal variation of the microclimate. However, the data collected has numerous outliers and error measurements. So I cleaned the dataset, filtered out outliers and extracted insights based on my analysis on clean datasets.

The Data

By taking a reading every five minutes from four different sensors: temperature, humidity, incident photosynthetically active solar radiation (PAR), and reflected PAR, the maximum number of readings we could have acquired is 50,540 real-world data points per mote. With 33 motes deployed into the tree, we could have recorded 1.7 million data points. The time span recording ranges from Tuesday, April 27th 2004 at 5:10pm and Thursday, June 10th 2004 at 2:00pm. There are two types of data recording, one through network and another through nodes. The combination of log file and network file gives us a reasonable dataset for the measurement period without losing any informaiton.

Data Collection

In the paper, the choice of measured parameters was driven by the biological requirements. We measured traditional climate variables – temperature, humidity, and light levels. Temperature and relative humidity feed directly into transpiration models for redwood forests.The platform also included a TAOS TSL2550 sensor to measure total solar radiation (300nm - 1000nm).

Data Cleaning

The paper found many anomalous data readings when examining the data. I processed network data and log data separately because the voltage distribution for both data sets differs.

1. Drop NA

The paper suggested removing all sensors that did not produce any readings at all, which included some sensors whose readings had been entirely removed by the voltage screening. The NAs always happen in pack of four which means none of the sensor works.

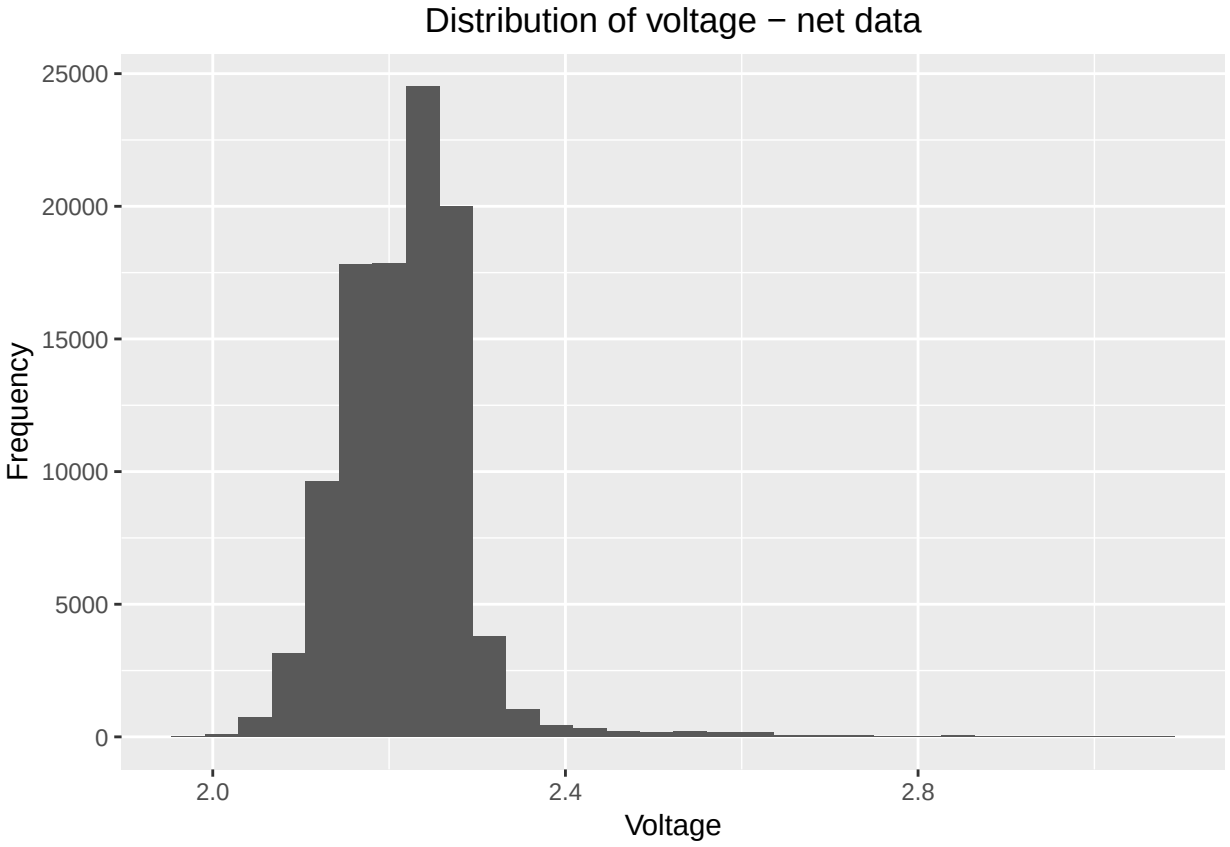
2. Exploration on voltage, humidity, temperature and adjusted humidity

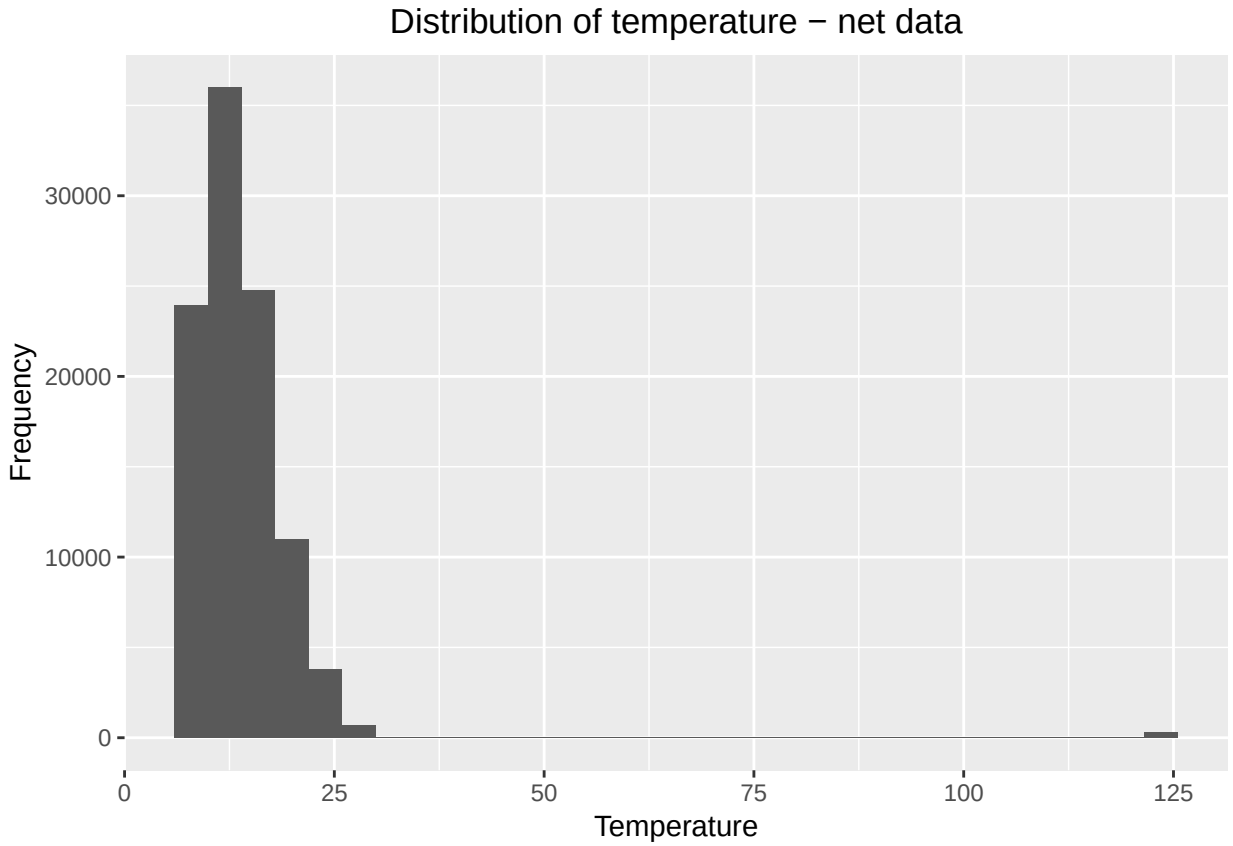
The paper found that battery failure was correlated with most of the outliers in the data. Once the battery voltage falls from a maximum of 3 volts to a minimum of about 2.4 volts, a node temperature reading begins to rise far out of the normal range. Other nodes that did not produce any correct data at all were also running on very low batteries. Prior to performing the analysis whose results are displayed above, we removed all readings that were taken when the node voltage was higher than 3 volts or lower than 2.4 volts. This removed nearly all of the outlying points in our dataset.

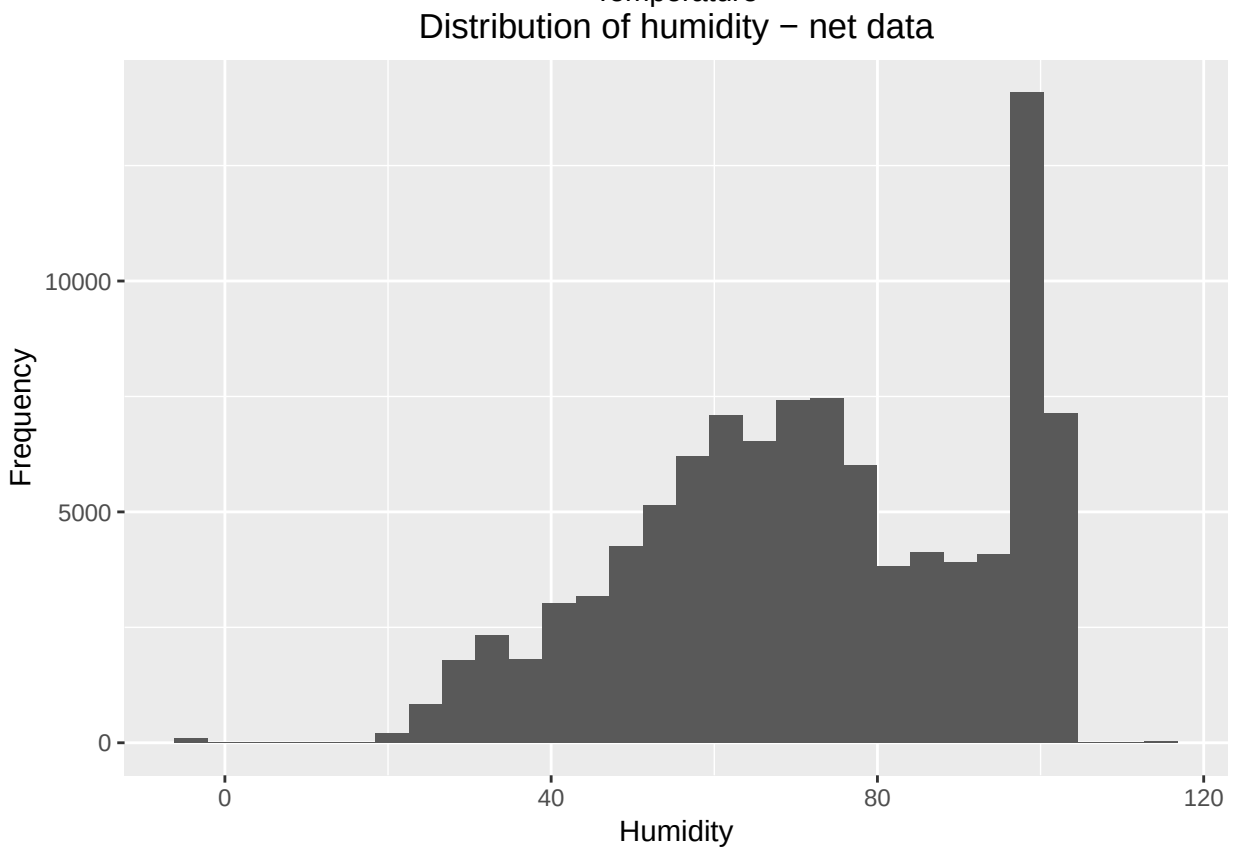
For net data set, I first plot the distribution histogram for voltage and filter out the outlier data points larger than 5. Then I plot the distribution histogram and boxplot for temperature and find that most of the temperature value falls in the range of 0 to 40, while small amount of the temperature data falls near 125. Then I plot the distribution histogram and boxplot for humidity. The paper also removed humidity readings that were above 100% RH but did not correspond to voltage problems. I also removed the humidity values that are below zero.

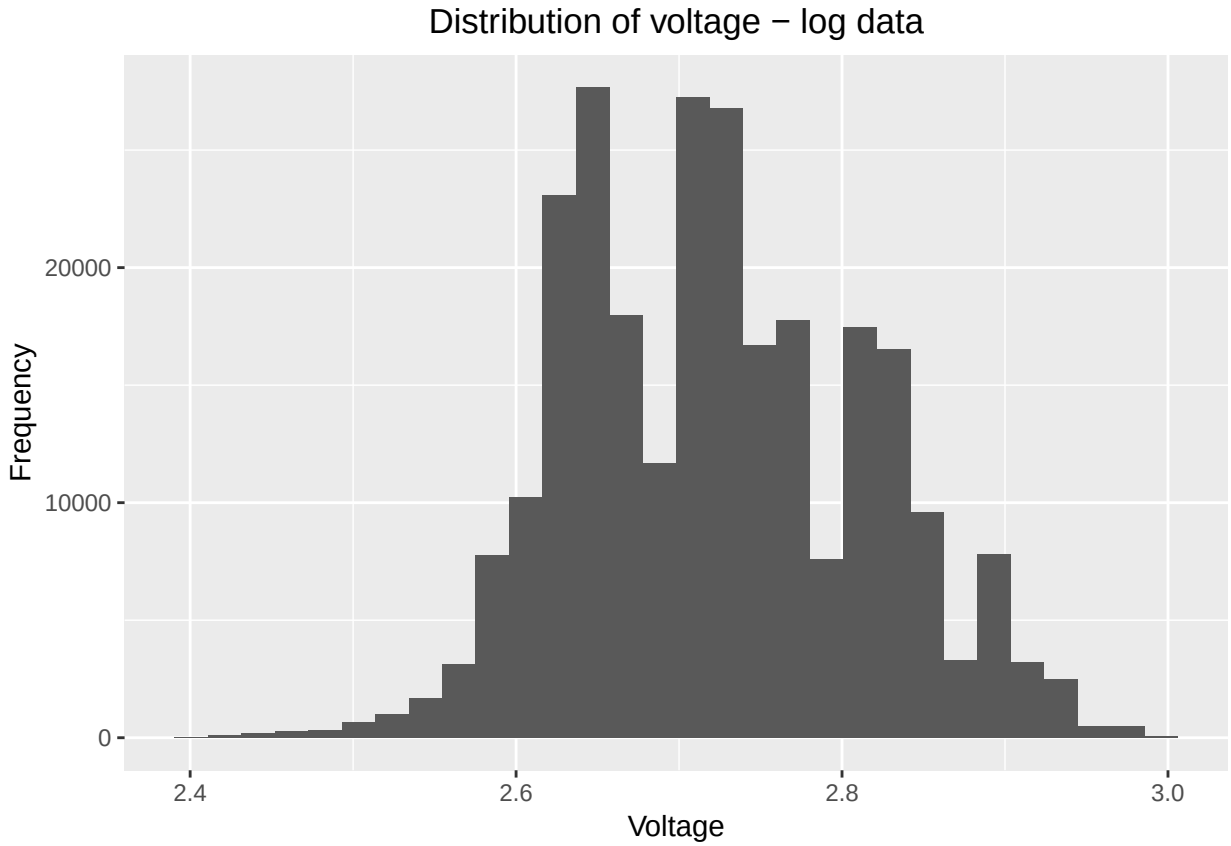
For log data set, similarly I subset the data where voltage falls in the range of 2.4 and 3.0. I then deleted humidity value that's lower than zero and higher than 100.

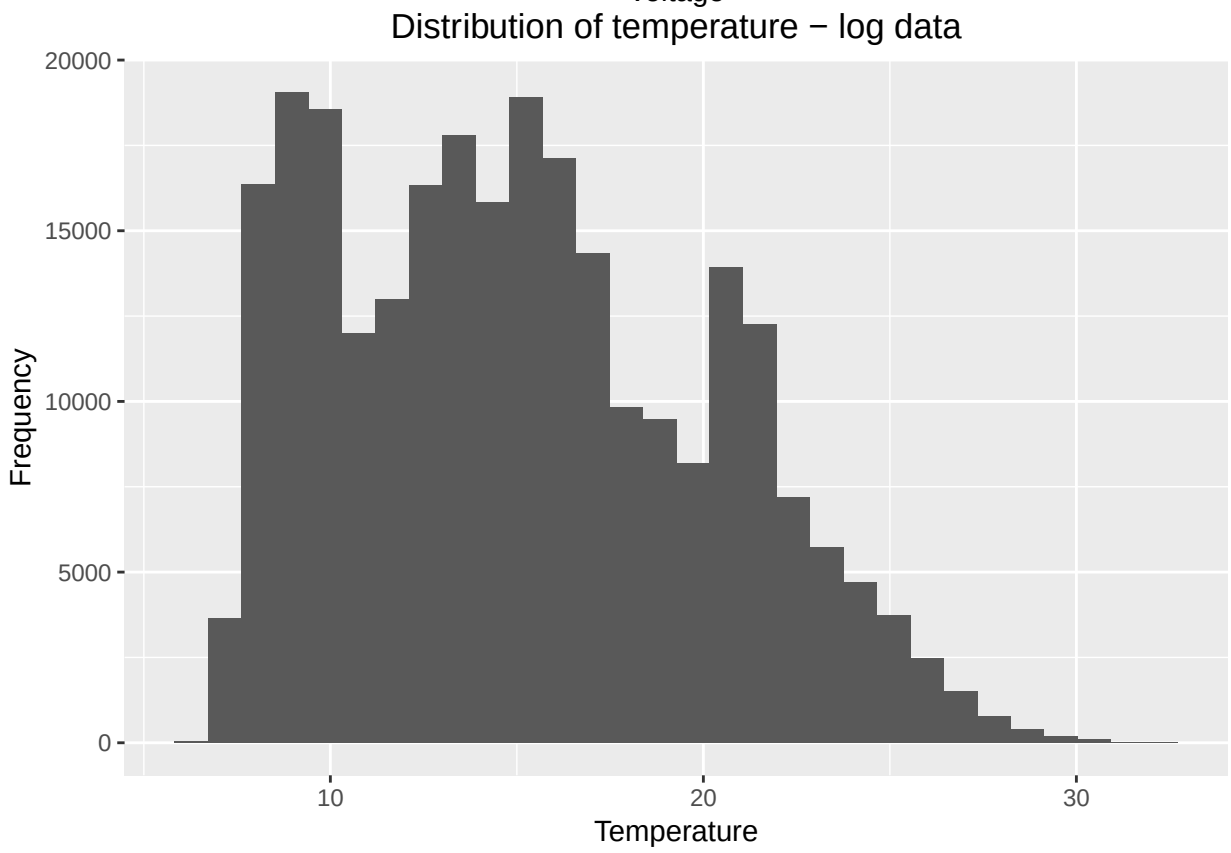
Then I plotted the pairwise relationship among humidity, temperature, and adjusted humidity for both net and log data sets.

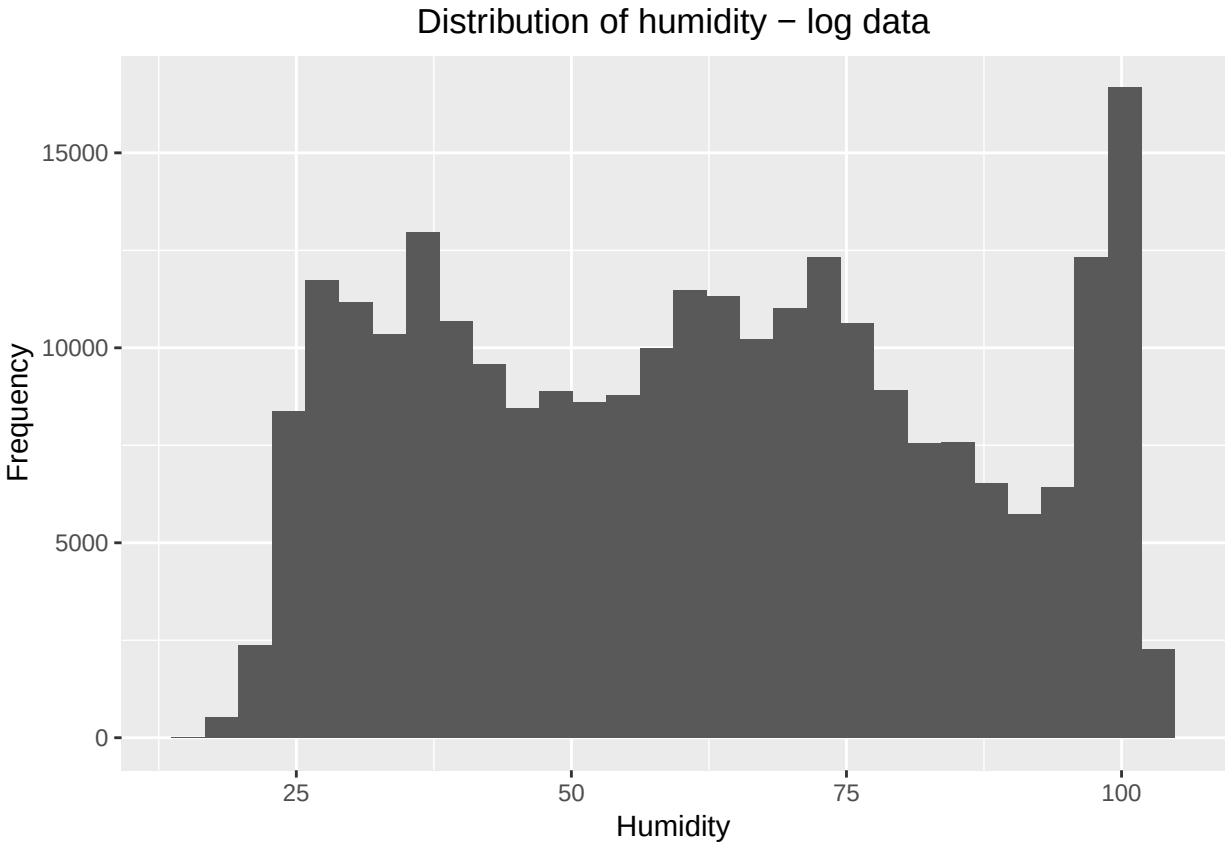



Distribution of temperature − net data
Frequency
30000
20000
10000
0
0
25
50
75
100
125
Temperature


Distribution of humidity − net data
Frequency
10000
5000
0
0
40
80
120
Humidity


Distribution of voltage − log data
Frequency
20000
10000
0
2.4
2.6
2.8
3.0
Voltage


Distribution of temperature − log data
Frequency
20000
15000
10000
5000
0
10
20
30
Temperature

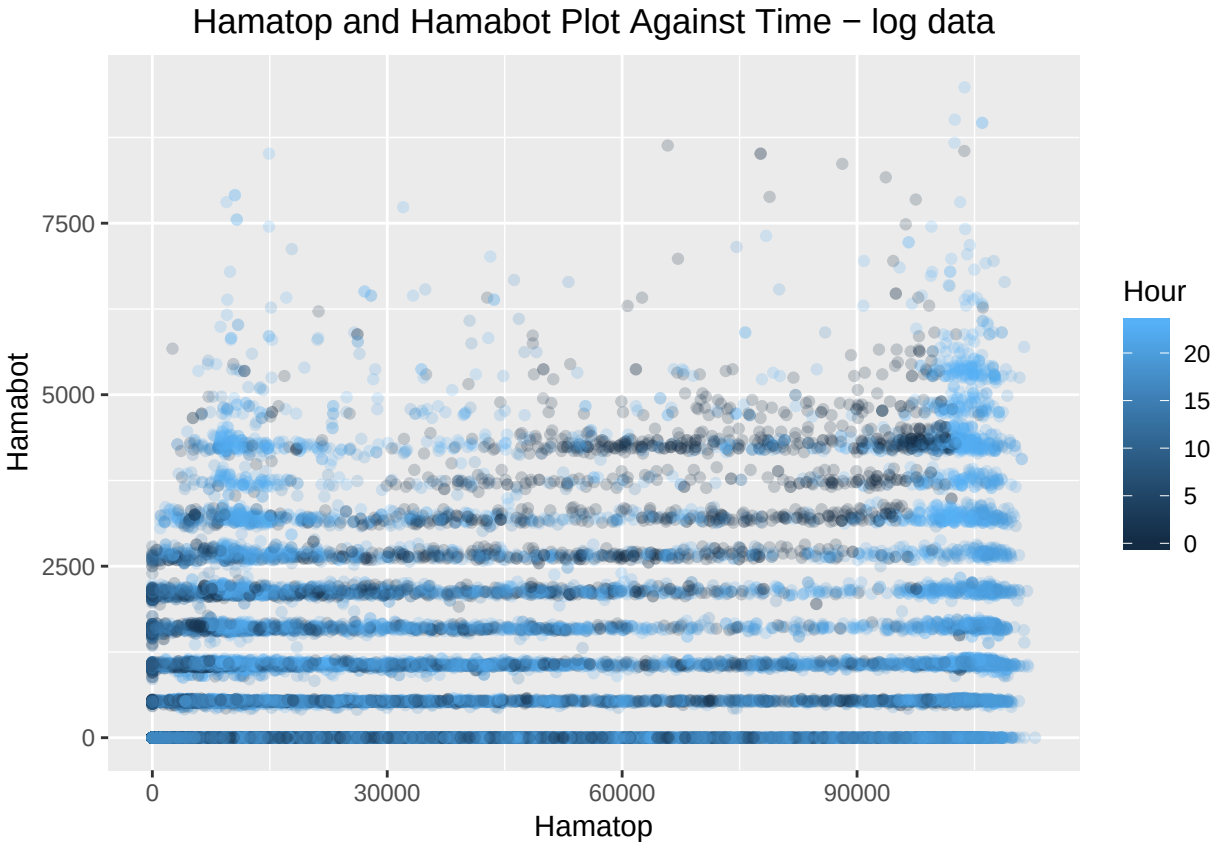


3. Exploration on hamatop and hamabot

I analyzed hamatop and hamabot on net data and log data separately. I observed similar pattern for both data sets. For hamatop, majority of the data falls in the range of 0 and 15,000 and there is one other small bump in the range of 90,000 and 120,000. For hamabot, most of the data falls near 0.

For net dataset, hamabot readings appear in unconsecutive value ranges. Also plot against time suggests that the very low and very high direct sunlight appears at late night, which doesn't make much sense. The result time column here is not accurate. In the later section, I decide to join the dates table and try using the datetime column there. Plot against humidity suggests that the very low and high direct sunlight appears in low humidity. The very low reflected sunlight appears in very low humidity as well.

For log dataset, hamatop readings higher than 150,000 are considered as outliers. I remove the hamatop value higher than 150,000 in this case.



4. Merge of two data sets:

I merged net data set and log data set datasets together, and then group by the combination of epoch and nodeid. Then for the repeated ones, I explored the data value with repeated epoch and nodeid combination, and then I decide to take the average value of each variable for each combination of epoch and nodeid combination.

We also want to merge the above table with dates and mote_location table. We want to use the datetime instead of the result_time for any time series plot, as the datetime is a more accurate recording of dates.

Data Exploration

1. Relationship between incident PAR and reflected PAR

I plotted incident PAR against reflected PAR and the coloring here represents the time difference from 12 o'clock noon in measurement of hour amount. I also plotted incident PAR against reflected PAR and the coloring here represents different level of humidity.

2. Time series measurement for single node

I normalize variable values including humidity, temperature, incident PAR and reflected PAR and plot the variable values against time for a particular node with node_id equals to 105. I also plot the variable values including humidity, temperature, incident PAR and reflectd PAR against time on a particular day (Apr 29 2004) for node with node_id 105.

3. Relationship between humidity and temperature against time

I plotted the relationship between humidity and temperature against time where the different colors represent different time stamp. I also set the filter on temperature here to be smaller than 40 in order to capture the temperature value in the range of [0,35] where majority of the temperature value falls into.

Graphical Critique

Critique the plots in Figures 3 & 4. What questions did they try to answer? Did they answer them successfully? Did they raise any questions not addressed in the text? Would you change them at all?

Figure 3(a) tried to understand the distribution of sensor readings projected onto each value dimension including temperature, relative humidity, incident PAR and reflected PAR. Temperature here shows a unimodal distribution, while humidity shows a bimodal distribution. In the incident PAR readings we see a bimodal distribution, while the reflected PAR readings do not show the same bimodal distribution.

Figure 3(b) shows the distribution of the sensor readings taken on each of the 44 days. The author wanted to find the temporal variables of the four sensors. The author wants to prove that by looking at the daily median readings, we can see weather movement in the large. However, the auther concluded that On May 7th, for example, we see that the bulk of the relative humidity readings lie above 95%RH, and at no time during the day is the relative humidity lower than 75%RH. This suggests that the entire tree is encased in fog, for nearly all of the day.

Figure 3(c) shows the distribution of all the readings taken by each sensor at different height. In the PAR readings, there is a very obvious spatial relationship. The author concluded that there is no spatial trend in the temperature and humidity readings. However, both the incident PAR and reflected PAR has negative value showed in the histogram and boxplot which is not consistent to the range analysis in the paper.

Figure 3(d) wants to answer the spatial trends of temperature, humidity, incident PAR and reflected PAR of the large microclimate system without temporal varaition.

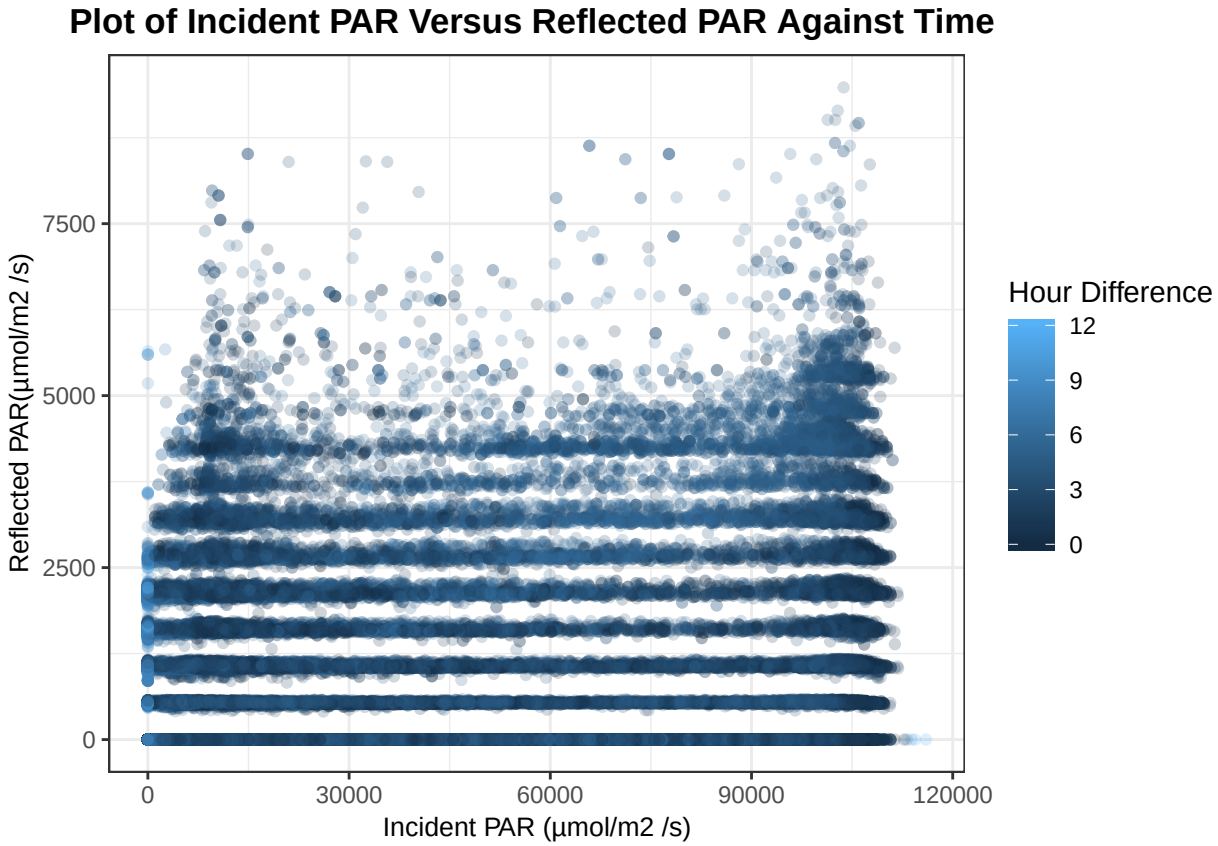
Figure 4 shows a day in the life of a redwood tree, as seen through network. The charts on the left top show the temporal trends of all the sensors, but they discard information about each sensor's location in space. The charts on the right place each sensor in its correct spatial location, and show the spatial gradients at a single moment in time. The dots in the left bottom chart represent individual light readings while the line represents the mean of the light readings at each timestep. The upper-left charts show the movement of temperature and relative humidity throughout the day. We see that both the temperature and the spread of temperatures throughout the tree increase as the sun rises. However, the graph didn't show whether the increased spread represents a particular trend over space or whether it is due to random local variations in air movement and solar access.

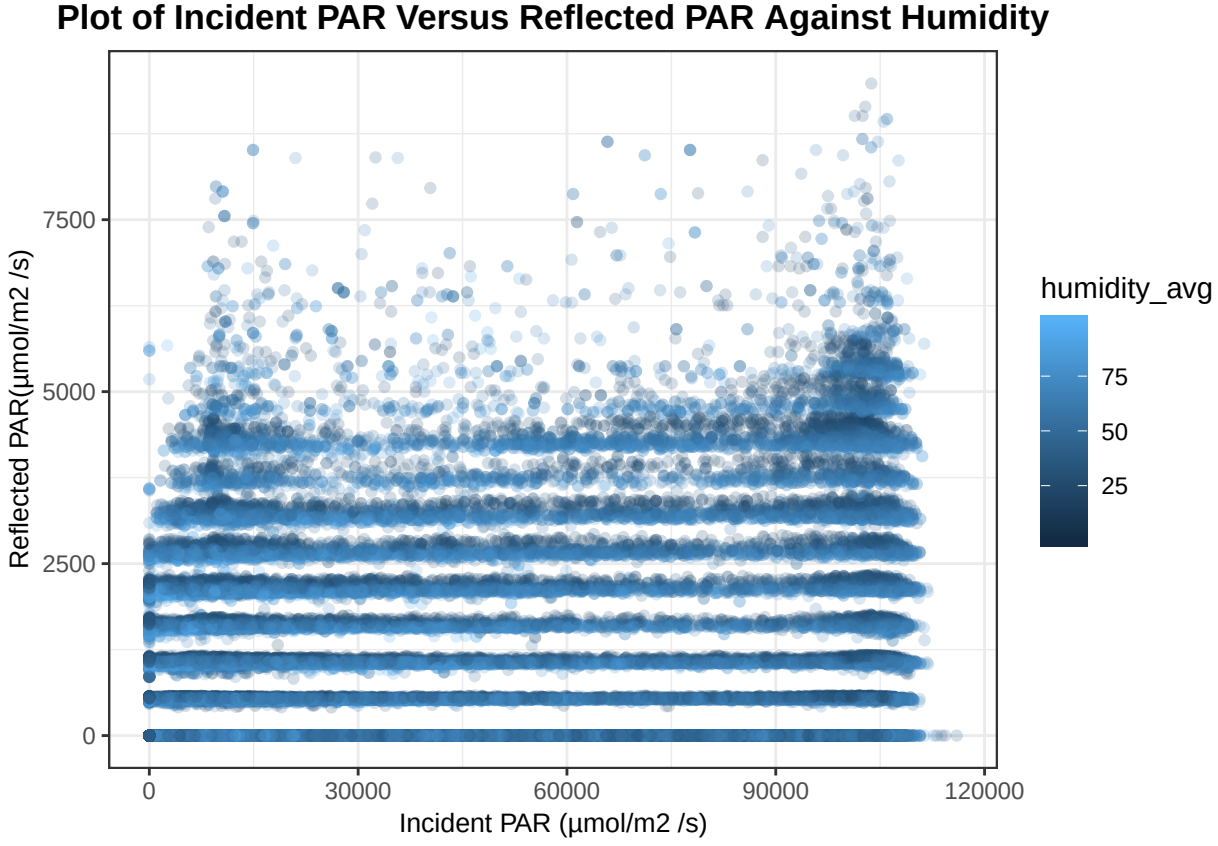
Findings

First finding

1. Relationship between incident PAR and reflected PAR against time and humidity

For each value of reflected PAR, there is almost continuous incident PAR value associated with it. Also there is no obvious pattern of PAR values against time, meaning that time doesn't have too much impact on the direct or reflected sunlight during a day. I also observed a relatively clear segmentation of humidity level associated with each reflected PAR range.


Plot of Incident PAR Versus Reflected PAR Against Time
Reflected PAR(µmol/m2 /s)
7500
5000
2500
0
0
30000
60000
90000
120000
Incident PAR (µmol/m2 /s)
Hour Difference
12
9
6
3
0

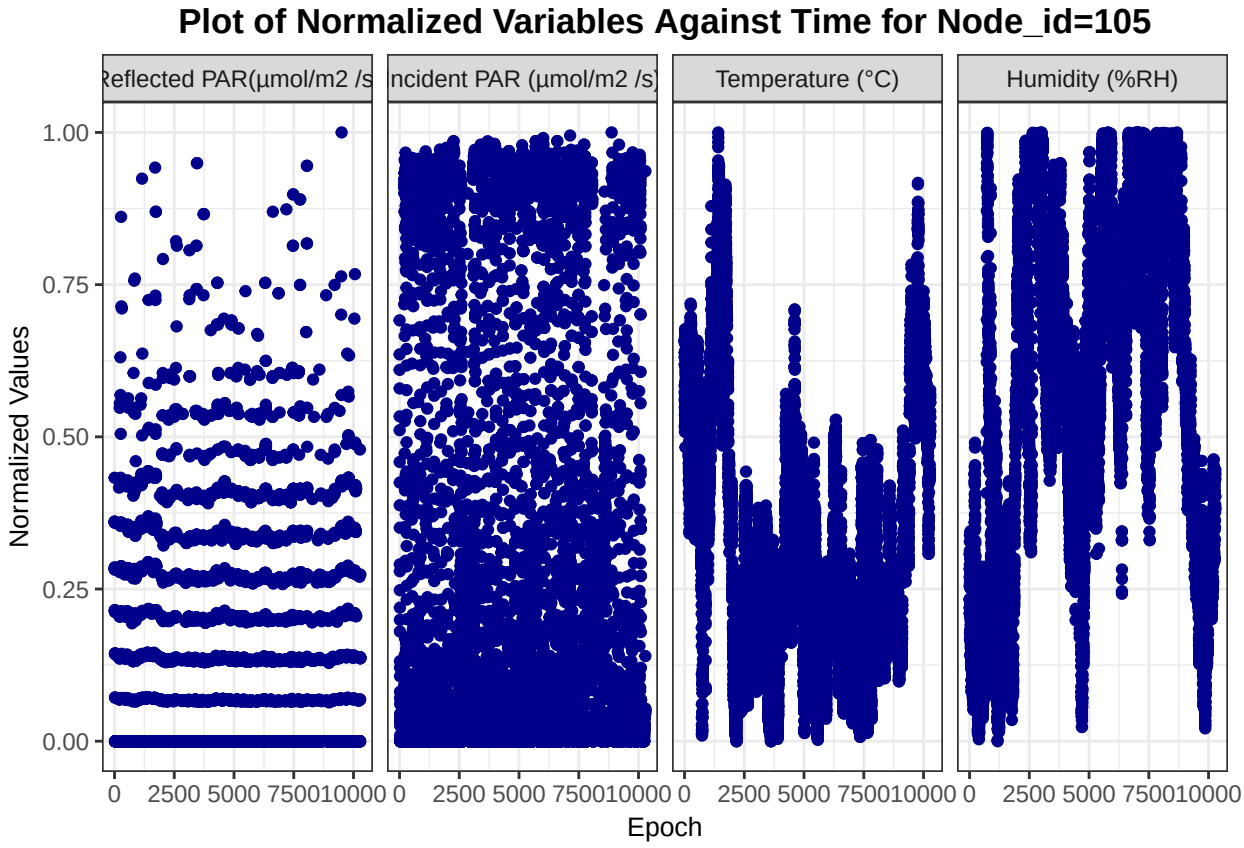

Plot of Incident PAR Versus Reflected PAR Against Humidity
Reflected PAR(µmol/m2 /s)
7500
5000
2500
0
0
30000
60000
90000
120000
Incident PAR (µmol/m2 /s)
humidity_avg
75
50
25

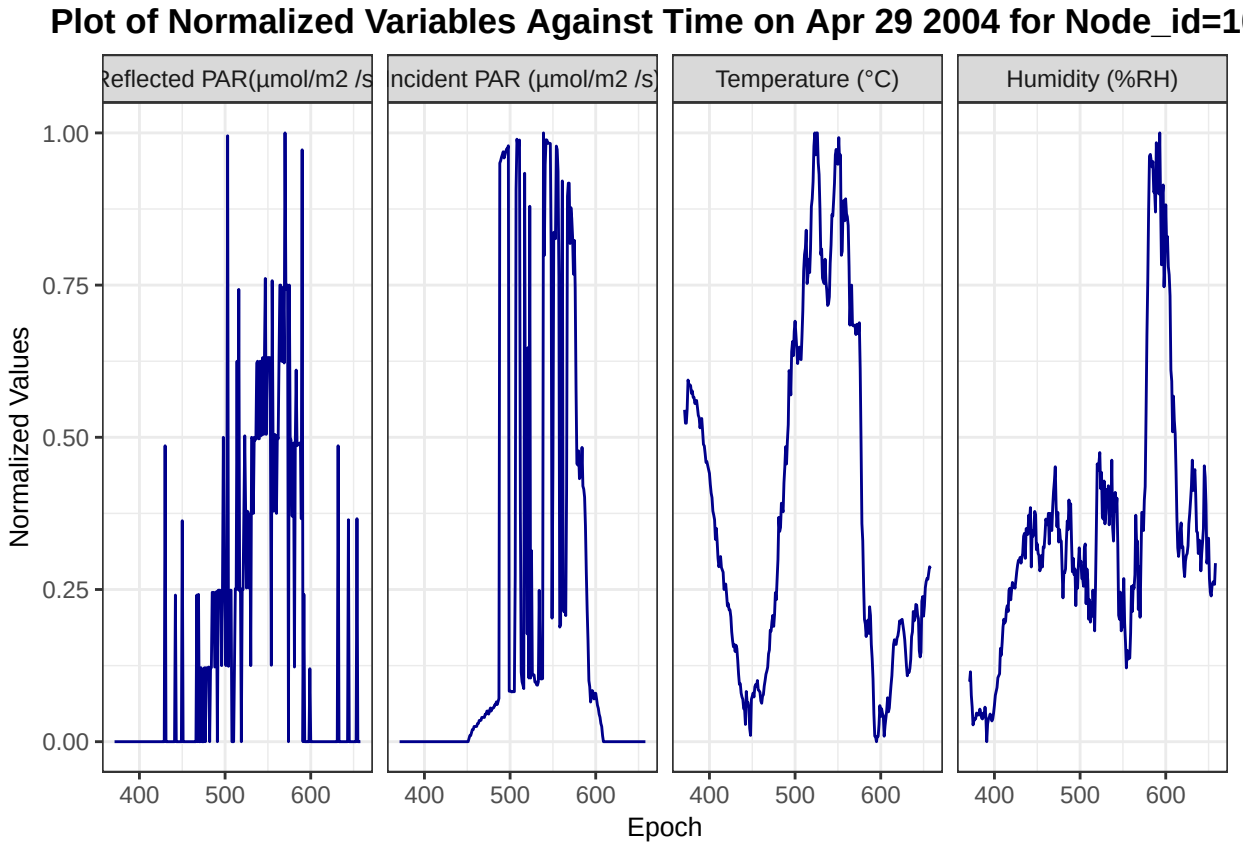
Second finding

2. Time series measurement for single node

I can tell from the graph that reflected PAR value only exists in certain nonconsecutive ranges and lower the reflected sunlight is, the readings are more stable with lower variance. Regarding incident PAR, there are more points concentrated in the lower and higher value range, suggeting exposure to direct sunlight is either very high or very low across time. Both temperature and humidity data have large variance across time.

We can observe from the graph that incident PAR and reflected PAR both surge during the day time and remains very low during night time. Temperature and humidity on a single day also have large variance.

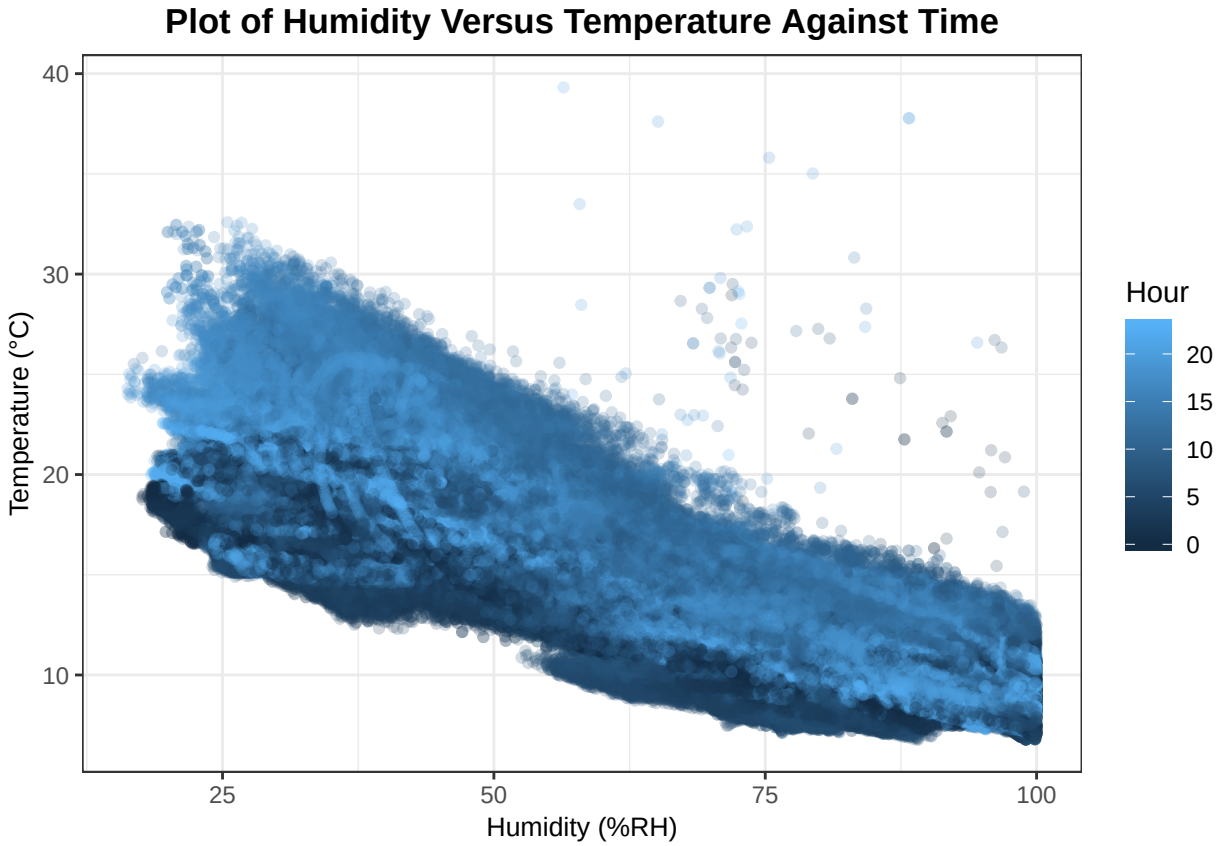




Third finding

3. Relationship between humidity and temperature against time

I obaserve a linear relationship between humidity and temperature: the lower the temperature, the higher the humidity. Across time, we observe that for each humidity level, the lower temperature happens from midnight to early morning time.



Conclusion

The data size is huge, but there are lots of error measurement. In this case, it is much harder to analyze temporal trends and spatial trends of each variable.

After cleaning and exploration on the data, I concluded below three findings on the variable measurement.

1. Relationship between incident PAR and reflected PAR against time and humidity

For each value of reflected PAR, there is almost continuous incident PAR value associated with it. Also there is no obvious pattern of PAR values against time, meaning that time doesn't have too much impact on the direct or reflected sunlight during a day. I also observed a relatively clear segmentation of humidity level associated with each reflected PAR range.

2. Time series measurement for single node (on a particular day)

I can tell from the graph that reflected PAR value only exists in certain nonconsecutive ranges and lower the reflected sunlight is, the readings are more stable with lower variance. Regarding incident PAR, there are more points concentrated in the lower and higher value range, suggeting exposure to direct sunlight is either very high or very low across time. Both temperature and humidity data have large variance across time. We can observe from the graph that incident PAR and reflected PAR both surge during the day time and remains very low during night time. Temperature and humidity on a single day also have large variance.

3. Relationship between humidity and temperature against time

Majority of temperature falls in the range of [0,35]. I obaserve a linear relationship between humidity and temperature: the lower the temperature, the higher the humidity. Across time, we observe that for each humidity level, the lower temperature happens from midnight to early morning time.