Redwood Data Analysis, Stat 154, Spring 2019

Yanran Li (SID:3034487587) Junho Woo (SID:23867648)

1 Data collection

1.1 Introduction

The macroscope allows us to collect a large amount of data that have the potential to advance the state of science by enabling dense temporal and spatial monitoring of large volumes. The purpose of this paper is to explore microclimate around a redwood tree by using wireless sensor network. To do this author and his teammate attached nodes with wireless sensor network around a Redwood tree. Each node measured air temperature, relative humidity, and photosynthetically active solar radiation. This wireless network captured a detailed picture of the complex spatial variation and temporal dynamics of the microclimate surrounding a redwood tree. Further, this study possibly can be used to validate other biological theories and study.

1.2 The data collection

The author and his team attached 80 nodes, mostly on west side, around a redwood tree in Sonoma California and recorded 44 days in the life of a 70-meter-tall redwood tree ~~in Sonoma California,~~ at a density of every 5 minutes in time and every 2 meters in space.

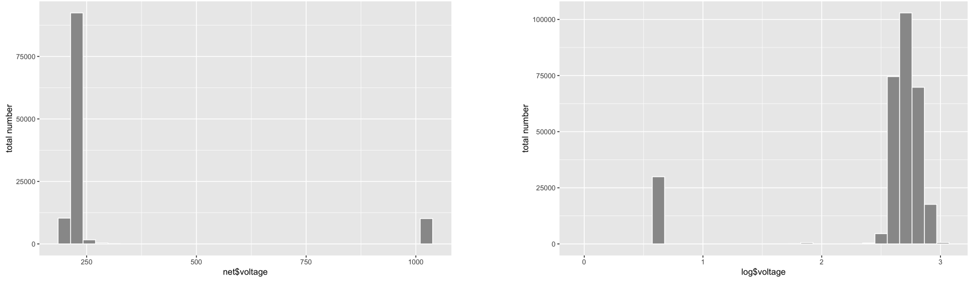
The data ~~were~~ collected from two systems, the TASK system and the local data logging system. The TASK system provides a query-based framework linking the sensor network to a database running on a gateway. Data were firstly stored in a local database to the gateway and then transmitted to another offsite database TinyDB. On the other hand, the local data logging system existed as a backup of network failure when no data ~~were~~ fetched from TinyDB. The data logger recorded every reading taken by the sensor on each query until the 512KB memory was full. The net dataset and the log dataset correspond to the two data collection systems respectively, so that they can complement each other.

To save the battery lifetime, each mote was awake for 4 seconds in every 5 minutes and collected the data. Main variables ~~that~~ we are interested in are air temperature, relative humidity, and photosynthetically active solar radiation, both direct one and reflected one. Sonoma-data-log.csv is the data for backups and a basis for analyzing the performance of the network whereas Sonoma-data-net.csv is the data that has been transmitted by the sensors around a redwood tree.

2 Data Cleaning

2.1 Variables discussions

First we checked histograms of each variables, including hamatop, hamabot, temperature, humidity and voltage measurements of both net and log datasets.One variable looked very suspicious and it was "voltage"(fig.1). Net data had voltage range from 198 to 1023 whereas



(a) voltage in net data

(b) voltage in log data

Figure 1: histograms of voltages in two data files

Log data had voltage range from 0.00906264 to 3.0302. Net data's voltage did not make sense, so we compared voltage of net and log data, and then do the voltage conversion.The conversion of the voltage measurement in the net to that in the log was calibrated by a simple linear regression with a Residual Square indicates 0.9960651 after removing a constant voltage reading of 1023 from node 134, 135, 141 and 145, and the graph is shown by Figure 2. Therefore we concluded that each voltage value of net data corresponds to each voltage value of log data. At the same time, the linear regression in R told us the voltage conversion is nearly $voltage_{log} = -0.01205 * voltage_{net} + 5.35135$. After replacement, we converted voltage from net data to have same voltage unit as log data.

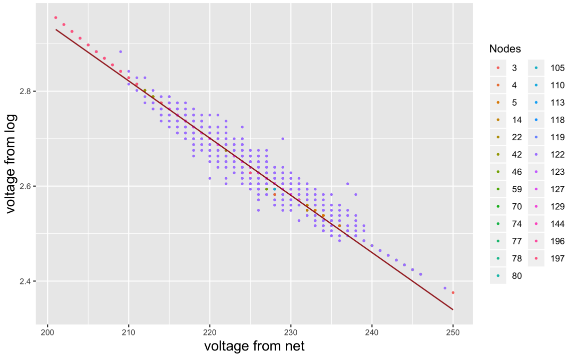


Figure 2: correlation of voltage between 2 data set

2.2 Remove missing data

To check if there is NA data, we used summary() function on Sonoma-data-all. We found that there were 12532 Missing value.There were 416036 total readings in "Sonama-data-all" and 12532 NA values were only 3.01% of the total, so we decided to remove this NA values.

What's more, we checked node ids that had missing values. Node ID 15, 122, and 128 had missing readings on variables that we need for analysis. Node ID 15, which was from log data, produced missing values from epoch 756 to 2435(corresponding time is from "2004-04-30 08:05" to "2004-05-06 04:00"). Node ID 122 worked in both log and net data and produced NA values from epoch 2806 to 9054 (corresponding time is from "2004-05-07 10:55" to "2004-05-29 10:37"). We found that not all values of Node ID 122 were NA; 9571 of 12371 readings were NAs. Finally, Node ID 128 had total of 2204 readings and 1366 values were NAs. Corresponding time period of those NAs was from epoch 845 to 2264, which was from "2004-04-30 15:30" to "2004-05-05 13:45". We used na.omit() function to remove just NA values.

2.3 Incorporate with location data

After using merge() function on the new "sonoma-data-all" with "mote-location-data", tree type information (edge or interior) was appended by a merge from the node location dataset and we transferred them into factors. Only the interior tree was the primary focus of the study and hence the data on the edge tree were all discarded. The data from nodes outside the deployment envelope (>1m from the trunk), with RH readings beyond the normal 0-100% range, or with negative adjusted humidity readings were also discarded. Now we have 15 columns in the combined data, and the number of variables are 12, which are "parent", "voltage", "depth", "humidity", "humid_temp", "humid_adj", "hamatop", "hamabot" , "Height", "Direc", "Dist", "Tree".

2.4 Outliers Rejection

In addition we used histogram and quantiles to visually identify easy outliers for the variables: humidity, humid temp, hamatop, hamabot. Also, we considered the range given on the original paper.

Starting with humidity, we noticed nodes 29, 78, 123, 141 and 198 were responsible for negative humidity readings.I discovered node29 produced constant humidity and constant temperature readings for the entire duration of the study, so I concluded it had a faulty humidity and temperature sensor. The hamabot/top readings did not go past epoch 600, which was too small to every verify periodicity, though the readings were within a reasonable range. I decided not to take a risk and deleted the (few and faulty) records of node 29. For the node 78, 123 and 141 with negative humidity, ~~they all had nothing was~~ within reasonable bounds (over 120 C temperature, highest hamatop or hamabot readings). Finally node198 simply had one outlier that produced a negative humidity reading (I checked the distributions by plotting all its other variable readings against time), so I deleted the one outlier.

Then we plotted the new data over time and it looked very reasonable and normal.

For temperature, the histogram indicated about 40 outliers. According to the range from the paper and quantiles, temperature should not be over 50. The outliers were from Node ID 3, 78, 123, 141 and 145. Comparing the above discussion about humidity, we concluded that the node 78, 123 and 141 would cause some problems. Then we threw all these Node ID with weird temperature away and did the ggplot, it looked normal and was inside the range.

We found that PAR unit should be converted. We divided the raw data by 54 so that we could match the unit with author's work.

For hamatop, there was one outrageous outlier, whose hamatop was over 15000 while everything else stayed within bounds. We found that the outlier was produced by nodeid 40, which basically went awry after some time. Recordings also last just 59 readings, so it was not a huge loss to delete these 59 data with extremely high hamatop(set this variable to NA).

For hamabot, 15% of the data don't fall in the range described by the authors for the reflected PAR i.e. 0 to 180. However, since we have a large dataset and it is feasible to do the analysis even after losing that, we trusted the author's recommendations and remove all the data with the reflected PAR lower than 0 and higher than 180(set this variable to NA). We would highly want the clarifications.

Just to be safe, I also looked at relationships between the variables. In order to keep most of the messages, I just set outliers' variables to NA and keep their other variables which may make senses, details can be found in our code. Only in this way we can get healthier analysis about all the variables in the following.

2.5 Other possible outliers

We then focused on the voltage and the plot told us that nodes nodes 3, 128, 134, 142, 143 had constant readings for the voltage (constant and close to zero). We investigated the other measurements taken by these nodes and plotted them against time, and found other variables were all acceptable. We then checked the network topology and checked neighbouring motes to see if the measurements were similar, and found no alarming difference, which confirmed my decision to keep their measurements, and only throw out their voltage recordings (set to NA).

3 Data Exploration

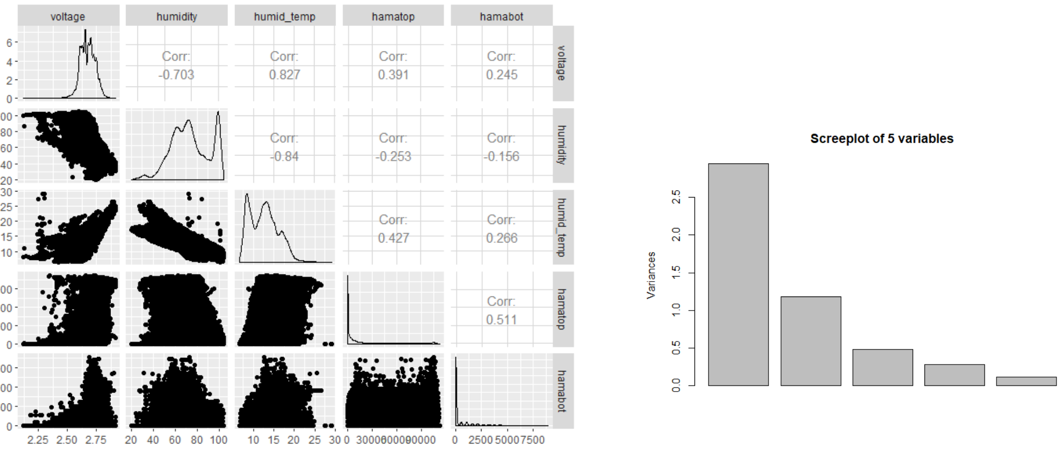
3.1 Pairwise scatterplots of variables

We picked the reasonable time period from epoch 3000 to 8000. We made histogram of epoch with our cleaned data and found that counts of epoch suddenly drops around epoch = 3000, and drops again around epoch = 8000. We divided into three time groups(0 3000,3000 8000,8000 end) and we used 3000 8000 because it had the longest time period of observation.

Then we made pairwise scatter plot on variables(see fig.3.a)(voltage, humidity, humid_temp, hamatop, and hamabot) to find the correlation between the variables. Pairwise scatter plot shows that humidity and temperature have clear negative correlation; the hotter temperature, the lower humidity. From the pairwise scatterplot, we could also see there was a strong positive correlation between temperature and voltage.

3.2 Predictors associated with Incident PAR

To find predictor of Incident PAR, We made scatterplots(with time range from epoch = 3000 to epoch = 8000), with 5 variables(humidity,temp,voltage, hamabot, hamabot). We used ggpairs function to see the correlation plot and the value. Incident PAR had the highest correlation(0.51067251) with Reflected PAR, and second highest correlation(0.42691844) with temperature. There was not much correlation between Incident PAR and voltage, and Incident PAR with humidity). This result makes senses because both Incident PAR and reflected PAR are the measurement of solar radiation so they should behave similarly. Temperature can be also a predictor of PAR because we can assume high temperature as sunny day and low



(a) pairwise of the 4 variables

(b) screeplots

Figure 3: Problem 3.1 3.2 3.4

temperature as cloudy day. Of course, PARs should be higher on sunny day than on cloudy day.

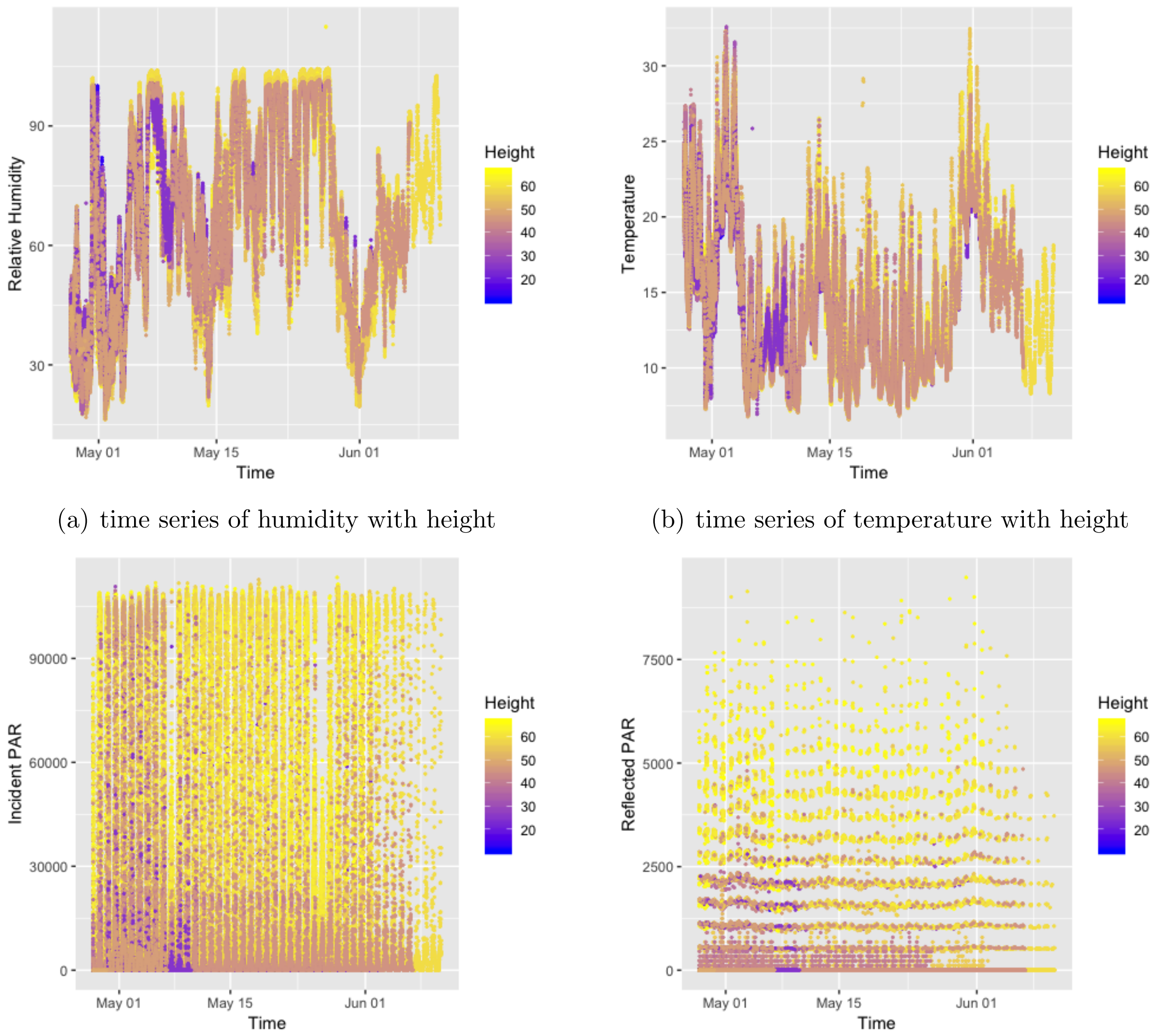
3.3 Temporal trend

First, we made this plot(fig.4) with all time period(epoch equivalent from April 27 to June 10) to see how height and variables are related over time period. We averaged temperature and humidity measured by all nodes. It is difficult to find some strong relationship among these pictures. But after May 8, we barely see the low height colors of nodes, it may explain some confusing.

The plots also show that both temperature and humidity looked continuous because the averaged temperature and humidity measured by all nodes did not change dramatically over time. Besides, temperature and humidity did not change dramatically over height, meaning that temperature and humidity measured in high height and low height did not have a big difference. Based on the Incident PAR(hamatop) and reflected PAR(hamabot) plot of height with color over time, we can see that the amount of PAR is related to the height. Our hypothesis was that the higher height would cause the more PAR, and these plots proved our hypothesis. However, there is discontinuity in PAR plots. This is because all sensor measuring PARs took 0 value at night or extremely cloudy day. Therefore, PARs plotting over time showed regular pattern of discontinuity at night time period.

3.4 PCA analysis

To do PCA, first we removed all the NA values from the cleaned data. Before doing PCA, we were worried about consistency of the data after removing NA value. So we checked the consistency after removing NA, we used summary() function to compare data consistency. We found that they were almost consistent(mean,quantile, min and max) even with 10% loss of data after NA removal. We have concluded that we could perform PCA by using the data after removing NAs. We then did PCA on the specific dataset; the dataset with only 5 variables(Voltage, Humidity, Temperature, Hamatop, and Hamabot). The scree plot(see



(a) time series of humidity with height

(b) time series of temperature with height

(c) time series of hamatop with height

(d) time series of hamabot with height

Figure 4: value, height and time

fig.3.b) showed that first two PCs have sd of 1.717 and 1.086, and other three PCs have eigenvalue less than 1. This means that we can summarize the data with some low-dimensional representation by using first two PCs.

4 Findings

4.1 First finding

Motivated in 3.3, we wanted to try different time scale to see the relationships among the variables. Since different days situation are not enough typical, we used the averages over days in order to observe the temporal movement of spatial gradients, by plotting time on the horizontal axis, height on the vertical axis, and the magnitude of the readings with color. The upper two plots of Figure 5 show the average movements of temperature and relative humidity throughout the day. In afternoon around 14-18, the temperatures are generally high while humidity stands in opposite. Also, the behaviors from lower part of the tree are usually followed by the upper part. The sunlight readings shown in the lower two plots of Figure 5 where both incident PAR and reflected PAR clearly show the movement of the sun over the period. As the sun rises at around 08:00, the top of the tree starts to get sunshine, and then the lower part experiences similar movement in readings until the sun sets at around 18:00, that is, the lower parts react slower than the higher. The plots show that the top usually led any movement in readings, which may because the buffering effect of the thick canopy on the lower west side. Hence, the upper part of the tree experienced larger variance in all microclimatic measures than the lower part within the same distance to the trunk. Thus, we can deduce that a node placed outside the deployment range, without the shield from the canopy, might have experienced larger variance.

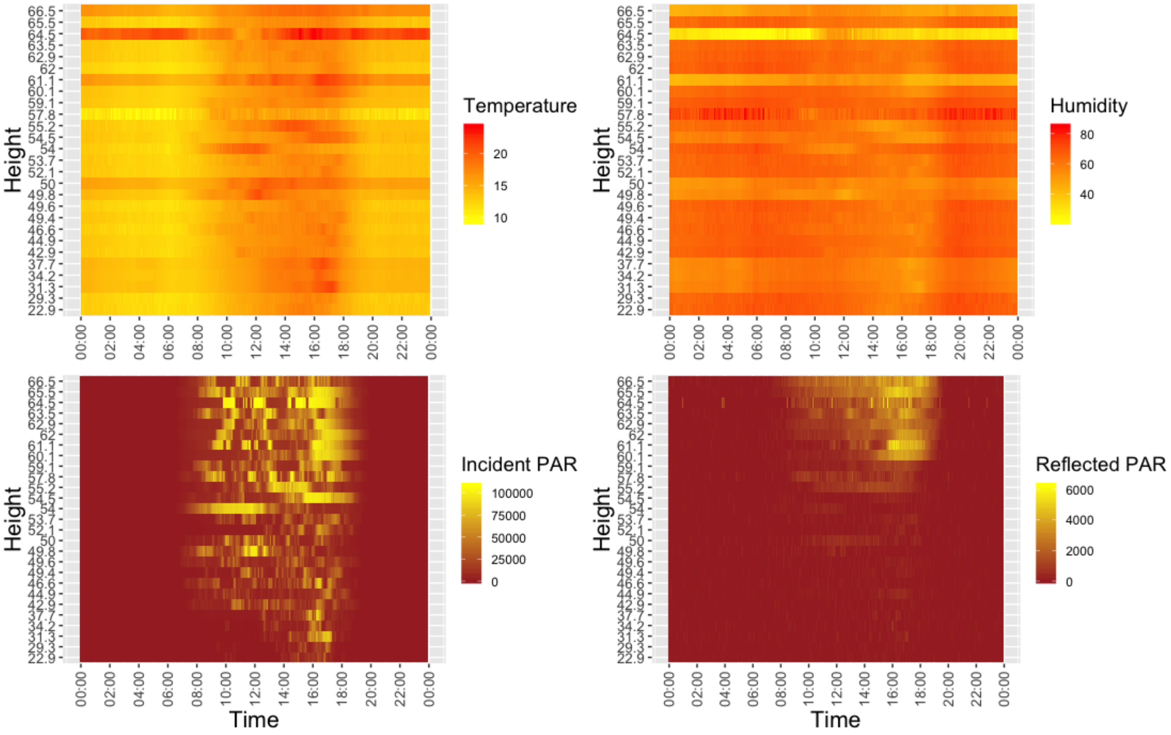


Figure 5: distributions about height with mean variables in day time

4.2 Second Finding

Our second finding studies the relationship of PAR readings with height, and colours by directions(see fig.6). Again, we restrict the epoch duration to 3000-8000 to ensure a constant distribution of heights of the motes. The box-plots(fig.), showing us that the outliers experiencing high direct PAR near the bottom of the tree were very sparse in comparison to those from the top. Colouring by the direction brings to our attention that the south-west side receive most PAR, which makes sence from our daily lives that the south-west side would receive more sunshine.

In both of the pictures, we can see that the lower locations means fewer outliers. We think it may because the higher part of the tree is easy to be disturbed by the experiment. Also, the hamatop has more outliers than the hamabot, which may because the reflection makes the PAR more stable.

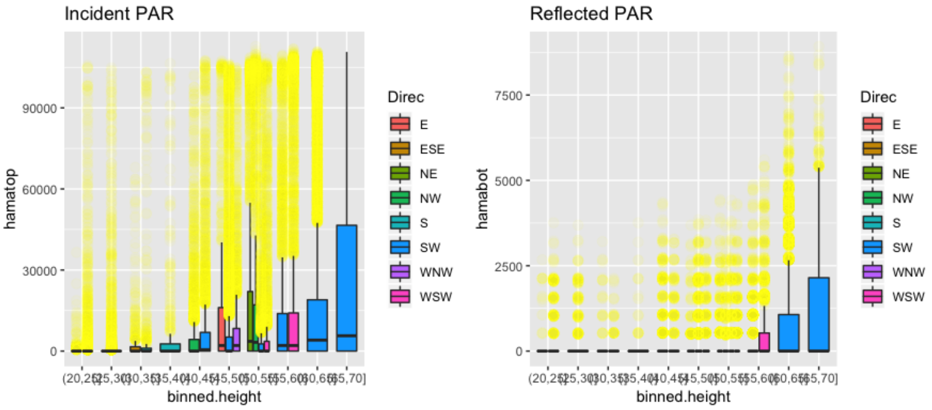


Figure 6: distributions about height with mean variables in day time

4.3 Third Finding

We wanted to explore deeper of 3.4. We divided time period into three groups (epoch 0 3000, 3000 8000, 8000 12000) and this was based on number of readings over epoch and at 3000 and 8000, we saw huge drop of number of readings, meaning that many nodes stopped working at epoch 3000(May 8) and ecoch 8000(May 25). We took PCA on these three different time period and compare each others(see fig.7). Surprisingly, we found that the directions of variables were different for different time period. Also, we could find that the voltage and temperature are pointing same direction whereas humidity are pointing almost opposite direction of voltage and temperature.These three variables' arrows(voltage, temperature and humidity) lie on horizontal axis. We found that hamatop and hamabot are pointing similar direction and look orthogonal to other variables. PARs arrows lie on vertical axis.

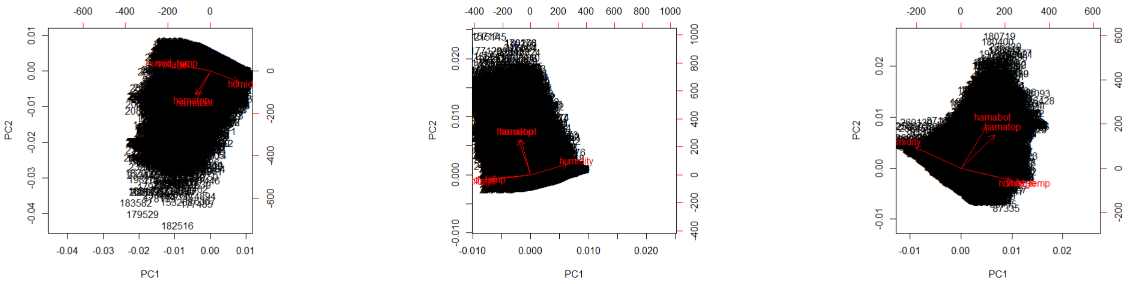


Figure 7: PCA analysis in 3 periods

5 Graph Critique in the paper

5.1 Figure 3[a]

In order to better plot figure 3(a), log transformation is appropriate because values on axis is too large and too wide(see fig.8). However, most values in PAR is zero and zero in PAR is meaningful so we can't just remove all 0s. In order to contain this meaningful 0, we added constant $c(=1)$ on PAR values so that all zero values become 1. Then we took the log transformation on the modified data. We made histogram with less breaks because large breaks made the author's histogram hard to read, and still less breaks convey conveys the message that there are absolutely large number of PAR 0 than PAR > 0.

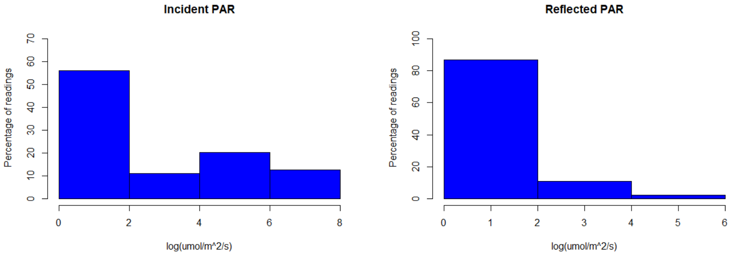


Figure 8: Histograms of Incident and Reflected PAR with log transformation

5.2 Figure 3[c] and 3[d]

The Figure 3(c) summarizes the distribution of each variable in a box plot over height level mixing temporal effects, and hence is not as informative as 3(d), which reduces time influence by centering the daily values for each variable. The PAR subplots clearly convey the message that lower-level nodes received less light than the higher ones, whether incident or reflected. However, the text hastily claimed that the lower nodes were colder than the higher ones on average, but there was no statistical test or obvious pattern in the temperature subplot. Also, this graph would have been more insightful if the intensity of color would have been used to show the amount and direction of the deflection of the features from the mean with respect to the height. In this way, we think our 2nd finding will be better(see fig.6), which convey more information that the author want to express. Also, motivated by fig4 in the paper, we think seeing the changes over daytime will do better to observe the temporal movement of spatial gradients. Using the averages over days will help a lot to discard specialities. What the author want to convey via Figure 3[c] and 3[d] is the relationship $height \times value \times time$. It's a 3D analysis so we generate new plots with the same data using the intensity of color(more details see in 4.1).

5.3 Figure 4

Figure 4 adds time and space information to the variable values on May 1st, when it was claimed to have the widest variation range for each variable. The left graphs show the temporal trends of all nodes throughout the day, and the right shows the spatial gradient at the moment when the corresponding variables changed drastically for all the nodes(we can see from the blue vertical lines on the left). The figure tries to present a 3D picture of the surrounding environment in a 2D setting. It is better than figure 3 because first two plots of figure 4 have their lines colored. However, there is no legend which color corresponds to which information; we assume the lines as different heights but still we have no idea which height represents which color. It would be better to divide height into thicker bin by averaging because there are too many thin lines that make us hard to see the pattern. Also it would be better to use both color and orientation of a triangle in the right panel to distinguish the nodes deployed on the side of the trunk.

5.4 Figure 7

Generally, plots in Figure 7 are not just great to obtain information. First plots tried to convey the message that majority net sensor could not perform perfectly and produced 0% yield, which composes of more than 35% of the total readings. Unlike net sensor, majority log sensor produced 50 80% of yield that composes about 60% of the total readings. To have better comparison, we suggest combining these two plots in one plot and make bars with two different colors so that we can compare productivity of net and log sensor directly with one plot. Second plots tried to convey the message of percentage yield over time. We could see that most of net sensor started working after May 5 and stopped working after June 2. Also, this boxplot show that net sensor produced around 40% of readings during this period. However, we think this boxplot is meaningless because we can see variance of percentage yield by nodes but we can't verify which nodes produce variance. The main point of this plot is to show the percentage yield of readings over time. We suggest that instead of using boxplot, we can just use histogram or line plot of average yields of readings by all nodes over time period. For the second plot of log, it is more meaningless than first one because this plot is hard to interpret. Again, instead of using boxplot, we suggest using histogram or line graph of average yield of readings by all nodes to show the percentage yield of readings over time. Then it would be easier to compare net and log. For third plots, instead of just blue points, we suggest putting node id next to y-axis values or next to the point. Also, making color different (like gradient) over height would show better visualization. In general, third plots are better than first two plots because we can see the meaning easily.

For the last plot of net, by looking at the plot, we could find that net sensor did not work from April 26 to May 6. This information corresponds to the second plot. One suggestion to make these last plots of net and log is to combine these two into one plot with three different colors. If we color differently on the lifetime of node in net, log, and overlapped period, we can easily check whether net or log, or both(overlapped) contributes to readings on certain date.

References

[1] Gilman Tolle, Neil Turner, Phil Buonadonna, et al., A Macroscope in the Redwoods, SenSys'05, November 2–4, 2005.

[2] We discussed with each other and consider every parts commonly. The leading member in each parts are as the following.

Junhoo: 1.1 1.2 2.2 3.1 3.2 3.4 4.3 5.1 5.4 Yanran: 2.1 2.3 2.4 2.5 3.3 4.1 4.2 5.2 5.3