STA 521: Project 1 Redwood Data

Yiliang Yuan & Yiyu Lin

2022-10-09

# 1 Data collection

## (a) Summary about the Paper

In the paper A Macroscope in the Redwoods, Tolle et al. describe how they get the previously unobtainable dataset of the microclimate surrounding a redwood tree through a wireless sensor network, and they show their analysis of these complex datasets. In the motivation part, Tolle et al. talked the purpose of this study is to research the ecophysiology of coastal redwood forests by using the wireless sensor network, from which researchers can show the capability of wireless sensor networks in advancing several scientific fields. All of the data was collected from a 70-meter tall redwood tree in Sonoma, California over 44 days. With a well-designed data collection process and detailed analysis, they finally reached the conclusion that the sensor network macroscope enables dense temporal and spatial monitoring of large volumes of data, which can contribute to the advancement of the scientific field. In this paper, Tolle et al. also talk about lessons they learned from the experiment and how they could deploy sensor networks better in the future. To help detect the failure of the network and make responses on time, Tolle et al. suggest that every long-term sensor network deployment should include a network monitoring component which can monitor the performance of the system and make alerts when the system behaves abnormally. Besides, by verifying the existence of the spatial gradients in the microclimate near the redwood tree and collecting enough data to view changes in these gradients over time, Tolle et al. were able to validate some biological theories related to these findings.

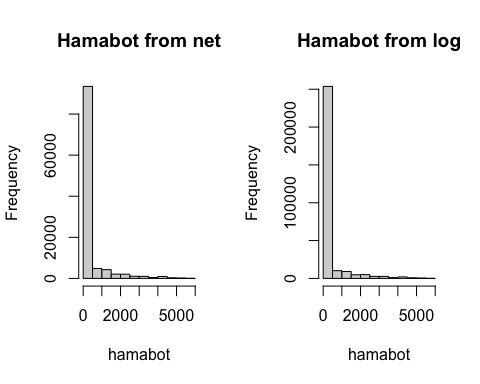
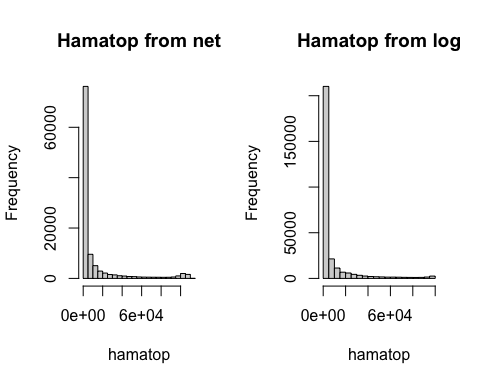
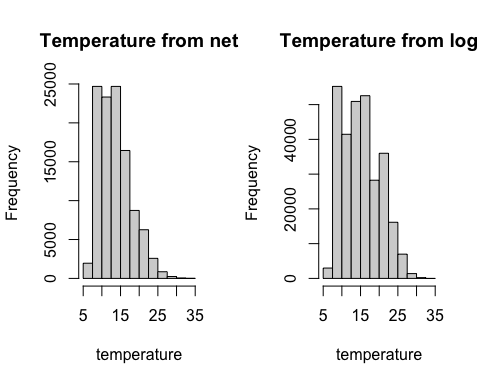
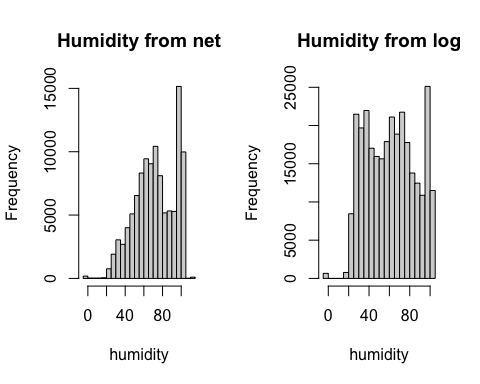
## (b) Summary about the Data Collection

The process of data collection is critical to this study, so Tolle et al. carefully designed the wireless sensor network system and used efficient deployment methodology for the system. Before being deployed, all the sensors were carefully tested and calibrated in two phrases which were roof and chamber. After that, sensor nodes were placed on the redwood tree 15 to 70 meters above the ground because most foliage was in the upper part of the tree. There was a 2 meters space between each node to ensure that Tolle et al. could capture gradients in enough detail. In addition, researchers placed the nodes on the west side of the tree since there was a thicker canopy and it could reduce the influence of the direct environmental effects. Besides, to record the microclimatic trends which had direct effects on the tree, Tolle et al. chose to place the node very close (0.1-1.0m) to the trunk. In order to get enough data, Tolle et al. record the data once every 5 minutes for 44 days.

In this study, Tolle et al. are interested in some variables including temperature, humidity, light levels and both incident and reflected Photosynthetically active radiation (PAR). After deploying nodes carefully, Tolle et al. collected data from both wireless sensor network and local loggers, so we have two different datasets which are sonoma-data-log.csv from local log and sonoma-data-net.csv from network. For the data received from the network, there are 114980 observations while the data stored in the local logger contained 301056 observations. The network data only has recordings from May 7th to June 2nd which was a period of 26 days, because most of the local nodes ran out of space on that day. In addition, data received from the network all show a voltage over 200, while data stored in the local logger show very small measurements of voltage.

# 2 Data cleaning

## (a)



Variables we are interested about are humidity, temperature, hamatop, which represents the level of incident PAR, and hamabot, which represents the level of Reflected PAR. We find that when scaled to the same range, hamatop and hamabot are consistent in the two datasets, but humidity and temperature are not consistent. In order to convert the data to the same range, we firstly see the range of those variables from the two datasets and we find that both data have many outliers which greatly influence the readability of the graph. Therefore, we decide to limit the range of each variable to where most data are clustered. We set the range of variables as: humidity (-5,120), temperature (5,35), hamatop (0,10^5) and hamabot (0,6000). After plotting histogram based on the range above, we find that graph of hamatop and hamabot are very close, but graphs of humidity and temperature are very different. Therefore, we could say humidity and temperature are inconsistent in these two datasets.

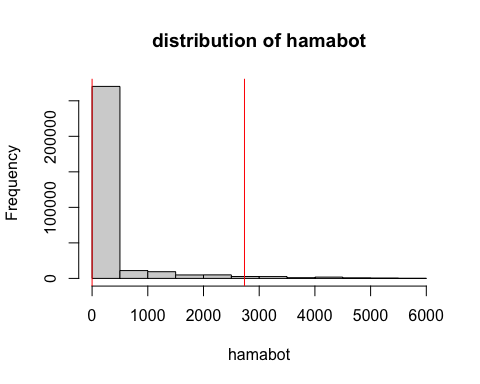
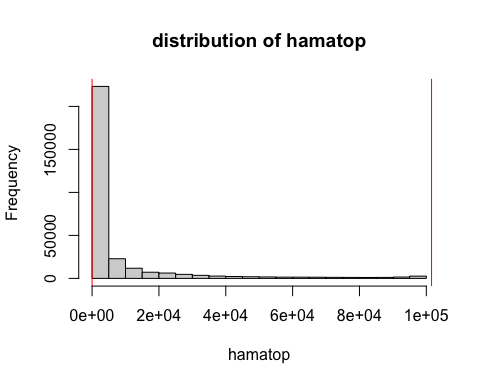
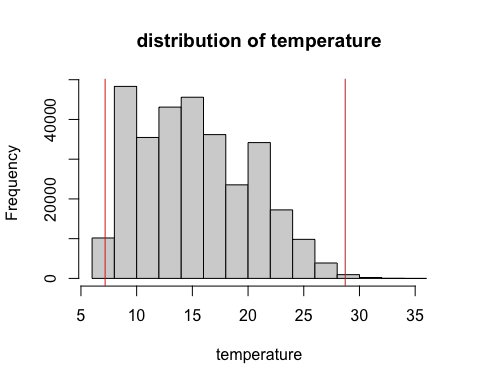
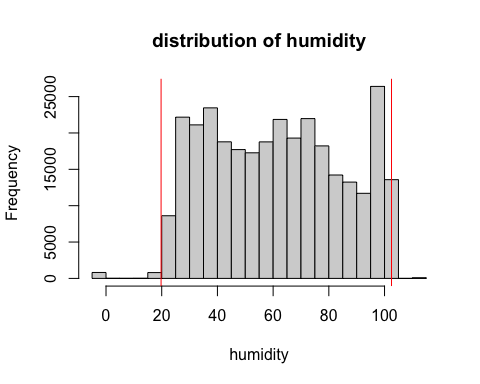
## (b)

We find that there are 4262 rows with missing values in the datasets from the net, and there are 8270 missing values in the datasets from the log. Compared to the total number of observations in the net and log datasets, we think that the number of observations with missing values only count to around 3% of total data, which is acceptable. For the data from net, we can see that all the missing values are from the node with ID 122, which means this node did not return information successfully sometimes. For the data from log, we can notice that except node 122, node 15 and node 128 also did not work well sometimes. For node 128, we find that its voltage is extremely low compared to other nodes, so this might be the reason why it did not work well. After finding the real date corresponding to the observations in data from log by its value of epoch, we notice that the time periods of data with missing values are distributed in 24 hours a day from April 30th to May 29th.

## (c)

We choose to use left join to incorporate mote-location-data.txt in the main table by matching the id of nodes. After the merge, there are 15 variables in the new datasets after the merge.

## (d)

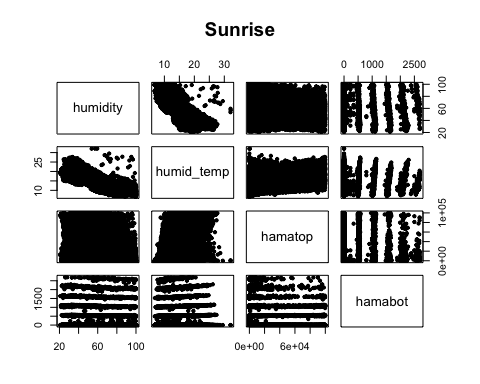
 On the graph showing above, we plot the histogram of these four variables with a 99% interval. For the humidity and temperature, we can roughly see a range in which most of the data are gathered. By visualizing the histograms of these two variables and considering researchers’ choices of the range for those variables, we decide to set objects with humidity outside of the range (15,100) and temperature outside of the range (5,33) are outliers. However, for hamatop and hamabot, which represent the level of incident PAR and reflected PAR, we find that there seem to be many outliers in these two variables and we should set a larger quantile to separate the outliers. For these two variables, we decide to use a 95% interval to roughly decide the data we want to keep. From the graph above, we can see that the range fro hamatop is around (0,10^5) and the range for hamabot is about (0,2700). Then, we can filter outliers with rough ranges we decided above.

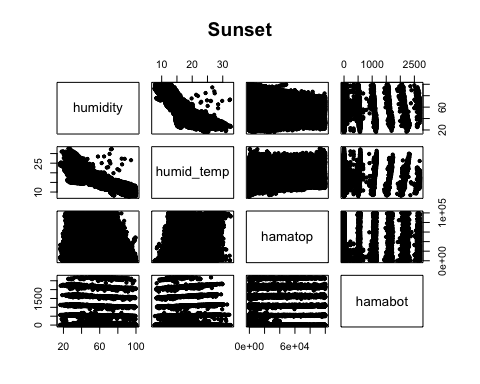
## (e)

When we are combining the datasets from both net and log, we find that there are some duplicated rows which should be deleted. Also, after reading the paper, we realize that data from node with voltage higher than 3.0v or lower than 2.4v should be removed, but we should keep those data from network with voltage all larger than 200.

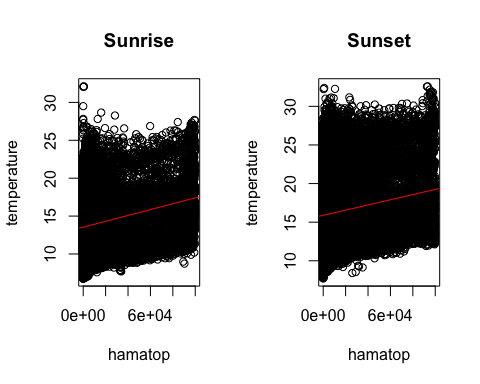
# 3 Data Exploration

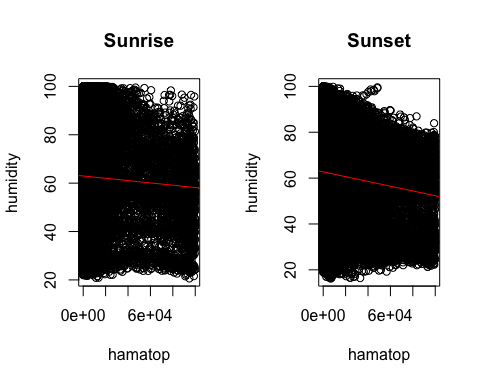
## (a)



 Before we get into other analysis on the data set after data cleaning, we should check the correlation among those variables we are interested in. We choose two distinct time period, sunrise (5:00:00 am to 10:00:00 am) and sunset (16:00:00” to 21:00:00) to check the correlation since these two time period can gives largest environmental change. The above two plots showed the pairwise scatterplot of humidity, temperature, hamatop, hamabot during sunrise and sunset. Both plots represents a clear negative correlation between humidity and temperature since most data points in humidity vs temperature plots seems gather along a line.

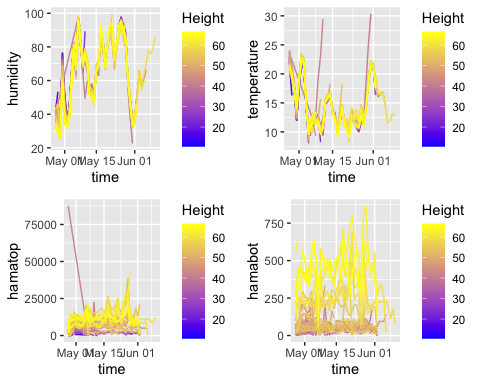
## (b)



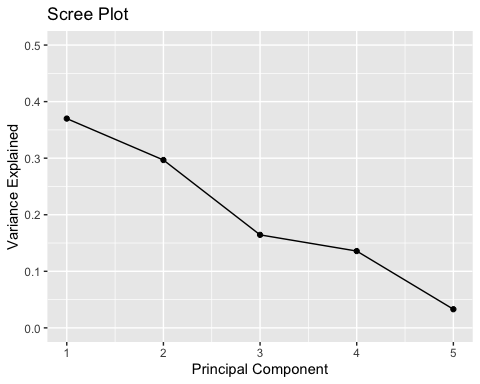


We plot the relationship between hamatop or incident PAR and both humidity and temperature. From the graphs above, we can see that there is a positive correlation between hamatop or incident PAR and temperature, and there is a negative correlation between incident PAR and humidity.

## (c)

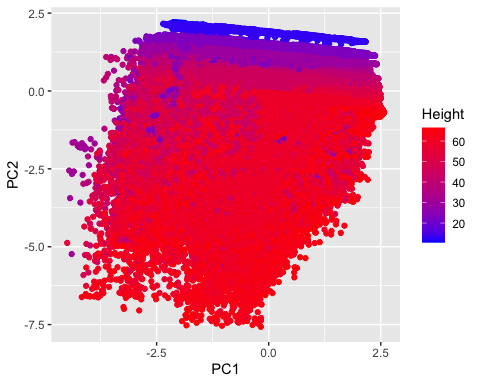
 The above four plots showed humidity, temperature, incident PAR (hamatop) and reflective PAR (hamabot), these four variables’ temporal trend. The temperature plot have the smallest range on temperature’s value, and if we ignore the strange outlier in temperature plot, the range could shrink more. Humidity plot have the second smallest range on humidity’s value, then the hamabot’s value range in the above hamabot (reflective PAR) vs time plot increase to around (0,825). For hamatop (incident PAR) vs time plot, since the existence of outlier, the range of hamatop’s value is expanded to the strange level, around(0, 87000). If we ignore this outlier, the range of hamatop’s value could be around (0,37500). For humidity and temperature, we could say there’s a continuity feature showed on the plot. The humidity tend to increase at the beginning of May, and decrease when the middle of May is approaching, then increase back to high level until the big drop at the end of May, beginning of June.The temperature tend to decrease at the beginning of May, and keep at a low level during the middle of May, then increase back to high level at the end of May. Incident PAR (hamatop) and reflective PAR (hamabot) seem keep at a similar level among the whole time period recorded in the above plots. This is strange behavior which doesn’t fit our expectation. There might be a relationship between these two variables and the time change, but the above plots don’t show clearly.

## (d)

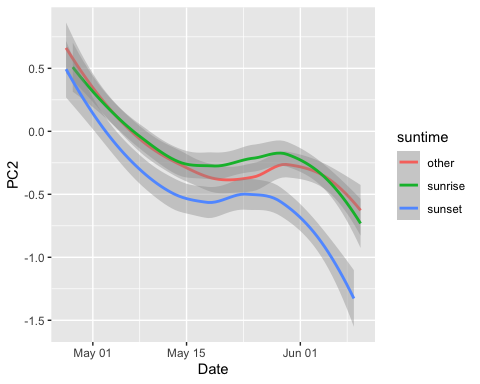
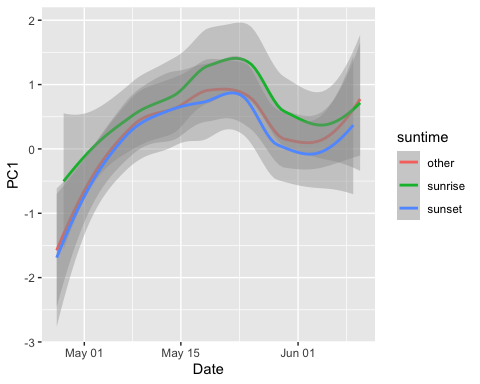
 We performed PCA with five dimensions including humidity, temperature, incident PAR(hamatop), reflective PAR (hamabot) and height. From Scree plot, we can see a clear elbow point when the number of principal component equal to 3. From the summary of principal component, we can see three principal components could explain around 83% the variation of the data points. Thus, we would choose the numebr of principal component to be 3.

# 4 Interesting Findings

## (a)

 We plotted the relationship between first two PCs with PC1 as x axis and PC2 as y axis, and we color our points with the value of heights. We find that there is an obvious stratification for our points based on the value of height, which is interesting. Points with smaller values of heights gather closely on the top of the graph, while points with larger values of heights group on the bottom of the graph. This might imply that there is a correlation between height and second PC.

## (b)

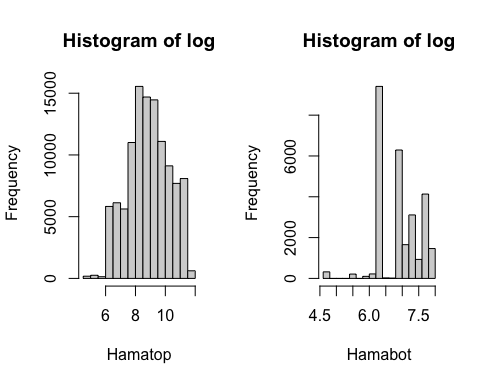


After getting new dimensions through PCA, we are interested in the relationship between our new dimensions and the time. We firstly separate our time period in a day to sunrise which is from 5:00 am to 10:00 am, sunset which is from 4:00 pm to 9:00 pm, and other which include all other time periods in a day. Then, we plot PC1 and PC2 with date as x-axis and draw the three lines with the time periods we separated before. We can see that the the trends in both graph are very similar, but we can notice the lines of sunrise look higher than the other two lines, and lines of sunset seem to be lower than the other two lines. Therefore, we find it is interesting that both PC1 and PC2 tend to be larger during sunrise period, while they tend to be smaller during sunset period.

# 5 Graph Critique in the paper

The overall quality of the paper by Tolle et al. is good. This paper gives us a chance to study collecting data and processing experiment through many graphs. However, some plots are not perfect from a statistician’s point of view. In the following content, we will discuss some graph critique in the paper. The focus will on the Figure 3[a], 3[c], 3[d], 4 and 7.

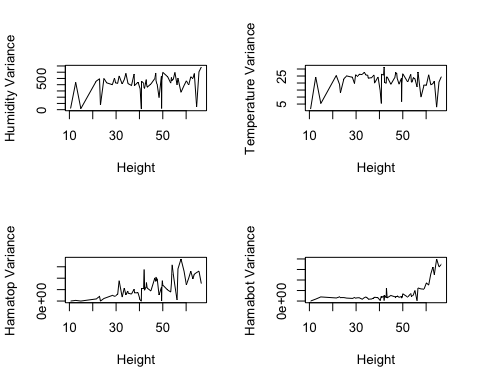
## (a)

In Figure 3[a], we notice the histogram of “Incident PAR” and “Reflected PAR”. Both these two graph have a tail of small-value or zero-value readings. Comparing to the large-value reading at the right of these two histograms, it would be very hard for readers to check the value of readings in the tail parts of these two histogram. In this case, I would recommend to apply an log-transformation on data and adjust the data scale. After the log-transformation, we could remove the zero-value from plots and shrink the value-difference between large-value reading and small-value reading. In this way, we could have a better visualization on the whole plot, especially for the histogram of Hamatop (Incident PAR). 

## (b)

In these two plots, Tolle et al. is interested in “temperature”, “relative humidity”, “reflected PAR” and “incident PAR”,these four variable’s variation at different heights. Tolle et al tries to apply the box plots to convey and visualize these messages, but this way doesn’t work very well.

Since in Figure3[c] and 3[d], there’s re total 27 different heights group along the y-axis. This means, in each plot, there will be total 27 box plots be included into each graph. If there is not a clear variation change along with the height change, which is the case here, it would be very hard for readers to tell the relationship between variable variation change and the height change.

From our opinion, Tolle et al. could calculate the variance at different height value and draw the line graph to check the distribution of all four variables vary at different height value.  From the four plots showed above, we could see there’s a clear increasing on variance of Hamatop and Hamabot, which means, with the height increase, these variation of these two variables (Hamatop and Hamabot) also increase. For other two variables (HUmidity and Temperature), the variation of them could not be tell from the above plots.

## (c)

For figure 4, it’s very hard to distinguish all the colors. It would be better for Tolle et al. to add the legend so that reader could understand the corresponding height of each color line. Moreover, the time period for these plot is 24 hours, but the temperature/humidity change period is not the whole day. If Tolle et al. could cut the whole day time period to several smaller time period, such as 5:00:00am- 10:00:00am (sunrise period), 10:00:00am-16:00:00, 16:00:00-21:00:00 (sunset period), and 21:00:00-5:00:00am. The visualization effect could be better.

## (d)

For Figure 7, it represents the relationship between yield and two predictor variables, height and day, recorded by network data and log data. This figure clearly showed both data sets include many missing data, since their “%” vs “day” box plots have a big gap. Moreover, for better visualization, we could apply a back-to-back histogram to compare the net work data and log data yield.