Lab 1

Yizhou Zhao, SID: 3032130362

September 14, 2017

1 Introduction

By following the instructions and ideas on the paper *A Macroscope in the Redwoods*, I came up with several questions which might be answered by the Redwoods dataset, which contains the space-time information about humility, temperature, incident PAR and reflected PAR. Within those dozens of sensors which were used to collect data, some of them sometimes were problematic. Therefore, a careful data prepossessing is necessary. After that, I would like to answer three questions in the finding part:

- Whether humitidy, temperature, incident PAR and reflected PAR data have much differences based on different spacial distribution?
- What the relationship between humidity and temperature at night ?
- Can the moonlight interpret the incident PAR and reflected PAR at night?

2 Data

2.1 Data collection

The 80 sensors were actually attach on two trees, with the heights from 15 meters to 70 meters. The data were collected in the early summer from April 27th 2004 to June 10th 2004, totally 44 days. Data were collected every 5 minutes, and the variables collected including *voltage, humidity, temperature, hamatop(Incident PAR), hamabot(Reflected PAR), and etc.* Also the spacial information of sensors were collected, including their locations, directions and distances to the tree.

2.2 Data cleaning

My data cleaning process can be divided into two parts: one for deduplicating the redundant data; another for removing problematic data and detecting outliers. My analysis and observations tell me that the *sonoma-data-all.csv*, which is the dataset we are interesed in, is the joined table from *sonama-data-log.csv* and *sonama-data-net.csv*. And a large number of its rows were duplicated. Besides, as the article mentions, some of the sensors may be out of function during the 44 days, and the extreme high or low voltage of the sensor usually leads to problematic data. Dectecting those outliers are also necessary in data cleaning.

Remove duplicates

First, let me put aside the problematic data with repect to 'voltage'. I assume that at each time, each sensor will report only one record. However, after checking the *sonama-data-log.csv*, I found that 23.3% of the rows are actually redundant, meaning that at that moment, one sensor has at least two different records. In most cases, those records on humidity, temperature, hamatop or hamabot are similar, except 26 cases, in which those records differ so much.(See more details in my cleaning codes.)

For those 26 cases, I deleted the records of them and treated those rows as missing values for analysis. For the duplicated similar records, I took the average for all the variables that we are interested in.

Remove problematic data

Starting with humidity data, I notice that somes of the nodes: 29, 78, 123 and etc. are giving me the netative values on humidity. After checking the records of those data, I found that sensor node 29 broke down from the very beginning of the experiment, showing constant values all the time. Then, I decided to delete all the records from node 29.

As for the other nodes giving negative values on humidity, I checked that they were doing well on other time points. I did not want to take a risking of delete all of their records, so I just delete those problematic ones.

Most nodes had a problem of wrong voltage during the experiments, which means that they had the voltage less than 2.4 or larger than 3 sometimes. As suggested in the paper, those data should be deleted. However, it is worthy mentioning that lots of voltage values are of hundreds level, such as 220 and 230. The researchers may collect those data while the nodes were charging. Personal experience tells me that electronic devices are likely to make mistakes while charging.

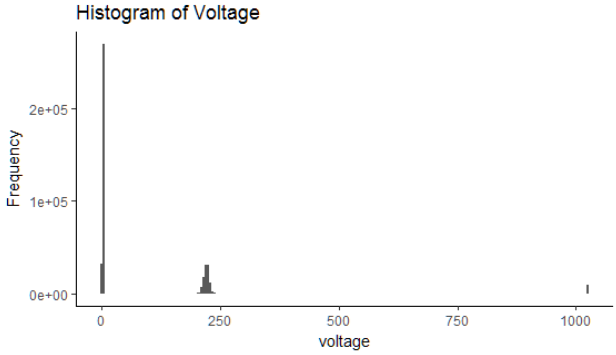

Figure 1: Histogram for Voltage data

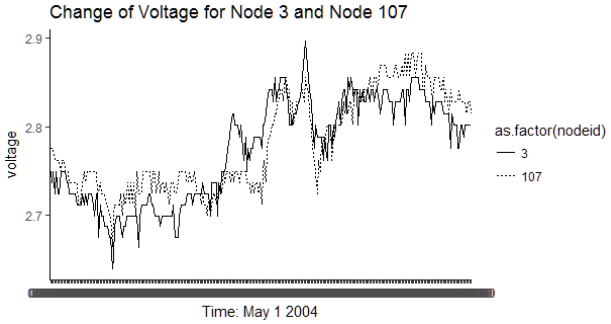

Figure 2: Two sample of the change of voltage

2.3 Data Exploration

The first plot (figure 3) is a histogram of how many observations were taken in each day. As we can most records were noted down at beginning of the experiment: from the end of April to early May. After May 8th 2004, the number of records suddenly went down. I do not know what happened and may be most diversified and high-quality data just came before that date. Therefore, in my finding part, I firstly would look at the data recorded at the beginning of May.

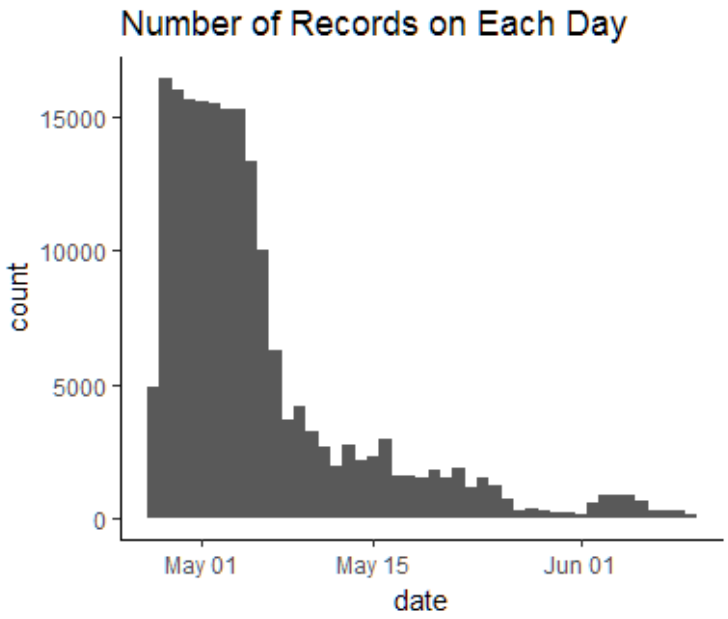

Figure 3: Histogram for records

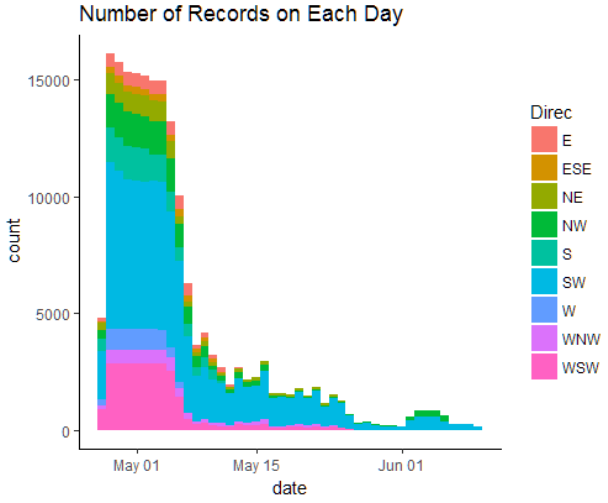

Figure 4: Histogram from differenct locations

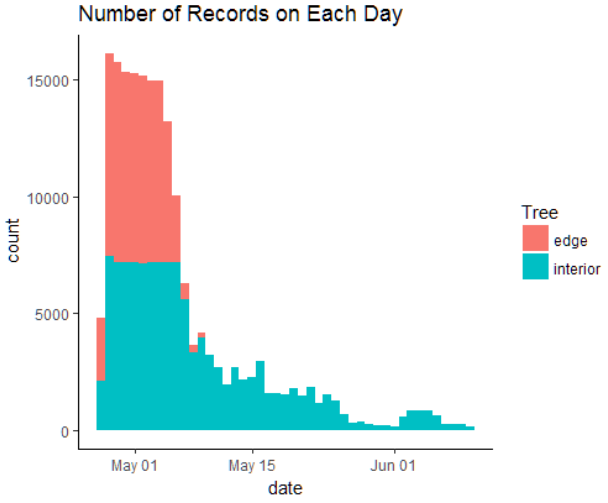

Figure 5: Histogram from differenct locations

Figure 4 and figure 5 tell the detailed information of where the records came from. Interestingly, the nodes located at the south west made the majority of the records. And the only the first two week recorded down the information from the edge of the three.

3 Findings

3.1 Finding one

In the first finding, I looked at the difference of humidity and temperature among different locations. As mentioned in the paper, there is probably some differences between different locations.

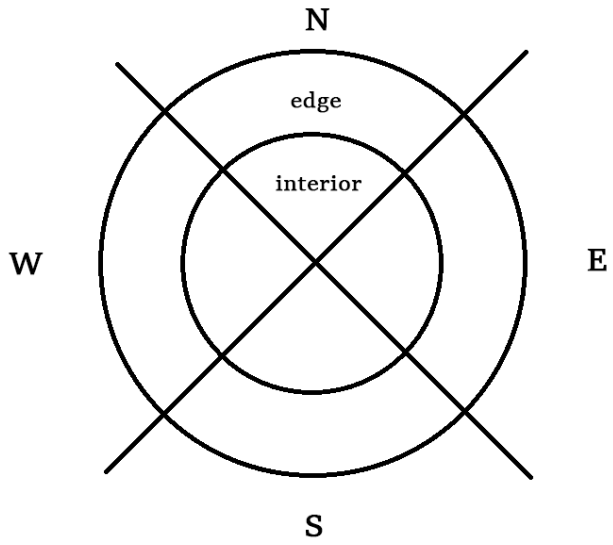

Figure 6: Locations of nodes

Figure 6 shows different locations in the study. After that, some careful aggregating process was applied to analysis the difference among them. I took the records at May 1st for example. Figure 7 shows the relationship between locations and humidity.

The horizontal line of x is the time, and the vertical line shows the value of humidity. As I could see from the plot, the directions of the nodes mater more on the interior and their differences were little if the nodes were located on the edge. The humidity usually reaches the climax at the middle of the day and the humidity at the direction of south, usually has larger values at noon.

4 Finding two

My second finding plots relationship between humidity and temperature at night. Their relationship in the day may be well analyzed as well. However, I would like to discover something special here and analyis the relationship during 0 o'clock to 6 o'clock. Again, we restrict the epoch during the early of May to give a meaningful result. To avoid confoun ding factors, I ignored the time issue and only looked the linear relationship between those variables. Actually, I saw a negative relationship.

5 Finding Three

The third quesion is like whether their is a relationship between moonlight and hamatop or hamabot at night. The answer is not. (Look at plots for more details.)

6 Discussion and Conclusion

In approaching large datasets in the future, I would spend more time on data processing. The dataset was large, there are lots of problematic data, so this was and I fear when I just do exploratory analysis that I am not being comprehensive enough, missing confounding factors.