STA 521 Project 1 Redwood Data Report

Yicheng Shen (Student ID: 2806571) & Yunhong Bao (Student ID: 2427527)

October 13, 2022

1 Data Collection

1.1 Background

With the advancement of technologies, humans are better equipped to collect, process, and analyze huge volumes of multi-dimensional data using well-designed hardware and sophisticated software. In order to fully understand and utilize large, multi-dimensional data, statisticians have developed applicable statistical tools for exploratory data analysis over the years. In this report, we present a detailed data cleaning and exploration process on a real data set — environmental data around a redwood tree collected by a group of biological and computer science researchers from University of California, Berkeley (Tolle et al. 2005).

The project led by Gilman Tolle in the early summer of 2004 was 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. The researchers were motivated by local biologists' interests in studying the ecophysiology of coastal redwood forests with modern technology and analysis techniques. Advised by biologists, the key objective of the project was to understand the microclimate over the volume of an entire redwood tree, including air temperature, relative humidity, and photosynthetically active solar radiation (PAR). In addition, each data point's node ID, time, and the sensor position were also recorded.

The researchers carefully examined the data through conducting initial multi-dimensional analysis and range analysis, visualizing notable temporal and spatial trends, performing a combined analysis of parameters of interest, and finally removing apparent outliers and calibrating values from sensors. They concluded their work by elaborating on their findings and experience. They emphasize the importance of installation and the necessity of a network monitoring component since tiny differences in positioning of nodes could cause large effects on the resulting data. Their study also verified the existence of spatial gradients in the microclimate around a redwood tree. Furthermore, their data could play a strong role in validating biological theories and building quantitative biological models on the sap flow rate.

Broadly speaking, this project, with the aid of interdisciplinary expertise, provides rich insights into the complex spatial variation and temporal dynamics of the microclimate surrounding a coastal redwood tree. More significantly, it illustrates the potential of wireless sensor networks to obtain large quantities of data and employs a multi-dimensional analysis methodology to reveal the characteristics of the microclimate, offering a valuable example that benefits future studies and usage of similar technologies.

1.2 Data Description

The field work of the project was conducted in a study area in Sonoma California. The data was collected by a sensor node platform consisting of a suite of small and intricate sensor nodes deployed in

various positions in the tree. The node operating system was run by the TinyOS and TASK software. Researchers chose to deploy sensor nodes at 15m from ground level to 70m from ground level, with roughly a 2-meter spacing between nodes. They were also deployed on the west side of the tree to minimize the direct environmental effects by a thicker canopy. For similar reasons, most of the nodes were placed very close to the trunk (from 0.1 to 1 meter). Several nodes were also placed outside of the interior tree to monitor the microclimate of the immediate vicinity.

The period of data collection process lasted for approximately 44 days. The first reading was taken on Tuesday, April 27th 2004, at 5:10pm, and the last one was taken on Thursday, June 10th 2004, at 2:00pm. With 33 motes deployed into the tree, researchers claimed that the maximum number of readings they could have acquired is 50,540 real-world data points per mote, with 1.7 million data points in total. Nevertheless, the available data in this report only exists from May 7th, 2004 to June 2nd, 2004.

The main variables of interest in the collected dataset are temperature, humidity, and light levels, or photosynthetically active solar radiation (PAR). Temperature is measured in degree Celsius ($^{\circ}C$), and humidity is measured in percentage of relative humidity (RH). The measurement of PAR, which suggests the energy available for photosynthesis, can be further categorized into incident and reflected PAR measurements. While the sensors initially recorded the PAR measurements in Lux, the researchers converted to the unit of PPFD (μmol $m^{-2}s^{-1}$) in their reported findings.

The data from each node in the mesh network were collected by a selection query using TinySQL and stored into `sonoma-data-net.csv`, a file with 114,980 rows of data. Researchers also extended their software architecture to include a local data logging system as a backup in case of network failure. The readings were passed into a flash log before taken by every query and eventually stored in the file named `sonoma-data-log.csv`, with 301,056 rows of data. In addition, the location information of each sensor node, or mote, was recorded in a separate document named `mote-location-data.txt`.

2 Data Cleaning

2.1 Unit Conversion

The first significant unit conversion is the conversion of voltage. We noticed the units of `voltage` are inconsistent between `sonoma-data-net.csv` and `sonoma-data-log.csv`. After performing a time series plot of `voltage`, `temperature`, and `humidity`, we were able to identify the battery failure time as depicted in the paper. However, utilizing the voltage unit from `sonoma-data-net.csv`, we see an explosion of observed voltage at failure times instead of a low voltage. Thus, it is determined that the real battery voltage in the unit of volts should be some inverse function of the voltage unit in `sonoma-data-net.csv`. Counting battery readings in both files, we observed a large repetition of voltage value 1023 and 0.580567 volts. Since the net data are a subset of log data, we determined that these two voltage readings are identical under certain transformation. The Analog-to-digital conversion (ADC) is employed here. We calculated the following conversion formula:

$$\frac{1023}{X} = \frac{ADC}{0.580567}$$

Voltage data in `sonoma-data-net.csv` is converted with formula. After conversion, all the voltage values matche readings in the log data file. Similarly significant, each value has a lower frequency in the net file than the log file, indicating the conversion is successful.

Unit conversion is also performed on variable `hamatop` and `hamabot`. After research, we discovered that

the PAR measurements are stored in the unit of Lux. Thus, we converted to the unit of PPFD (μmol $m^{-2}s^{-1}$) by a conversion factor of 54 — the conversion coefficient of sunlight since its the primary light source.

2.2 Time Restoration and Missing Data Removal

Another crucial element of the data cleaning process is the restoration of time variable. In `sonoma-data-log.csv`, the time variable is fixed at 2004-11-10 14:25:00, which is the log data extraction time. For data analysis purpose, we want to restore the time variable to actual collection time. This is accomplished by evaluating the `nodeid` and `epoch` variables. It is observed that `nodeid` represent a specific sensor and `epoch` records the exact number of time it records data. Furthermore, we discovered that `epoch` is synchronized across all sensors. In other words, an identical epoch reading indicates two data are collected roughly at the same time. Using this information, we are able to select a epoch as a benchmark and trace back and forth by adding or subtracting a calculation of time. We utilized data point `result_time` = 2004-05-07 18:24:58, `epoch`= 2812 from `sonoma-data-net.csv` as a benchmark. Then, we used the fact that an increase in epoch by 1 represents a five minutes time lapse to formulate the following time restoration formula:

$$\text{result time} = \text{ 2004-05-07 18:24:58 } + (\text{epoch} - 2812) \times 5\text{min}$$

Utilizing this function, the data collection time in file `sonoma-data-log.csv` is restored. Utilizing the time variable in `sonoma-data-net.csv` as a reference, the calculated time matches with the real time data by a minor error.

Furthermore, we noticed that the time variable in `sonoma-data-net.csv` is also suspiciously inaccurate, with higher `temperature` recorded during nights. We thus matched the `epoch` and time variable with those provided in Tolle et al. (2005). Based on the match, we adjusted the time variable by approximately seven hours.

After successful restoration of time, we start the process of removing missing data. In `sonoma-data-net.csv`, 4262 data points are missing throughout May 7th to May 29th. In `sonoma-data-log.csv`, 8270 data points are missing throughout the April 30th to May 26th. Data points with missing values are removed.

2.3 Data Merging

After unit conversion, data in `sonoma-data-net.csv` and `sonoma-data-log.csv` are ready to be merged into a final data file. To perform this task, we first investigated repetitive entries within each data file. An intriguing finding was that there existed data points sharing the same node id and epoch value, ie, certain sensors recorded two readings at a single time. This finding provided guidance to the method we should employ in the data-merging process. To avoid repetitive data, a row bind is performed on the mutated `sonoma-data-net.csv` and `sonoma-data-log.csv` file. Then, we selected distinct elements by the nodeid and epoch column. Through this method, only the first datapoints with repetitive nodeid and epoch values are kept. Repetition is avoided. Then, we performed a left join to combine the location data.

2.4 Outlier Rejection

Upon a close examination of the histograms of variables of interest, we noticed that the PAR measurements for sensor node 40, although having a seemingly normal battery voltage, were unstable and quite

abnormal as seen in Figure 1. Therefore, we removed the readings recorded by this node.

Figure 1: The anomaly readings recorded by node 40.

Moreover, the investigations by Tolle et al. (2005) suggest that low or malfunctioning batteries are the main cause of anomalous readings. Specifically, the researchers found that once the battery voltage falls from a maximum of 3 volts to a minimum of about 2.4 volts, a node's reading begins to rise far out of the normal range. Therefore, we removed those readings taken when sensor's battery was not in the proper range. In addition, we also removed the data that reported a `humidity` reading over 100%RH. Figure 2 presents the comparison of before and after removing anomalous sensor readings.

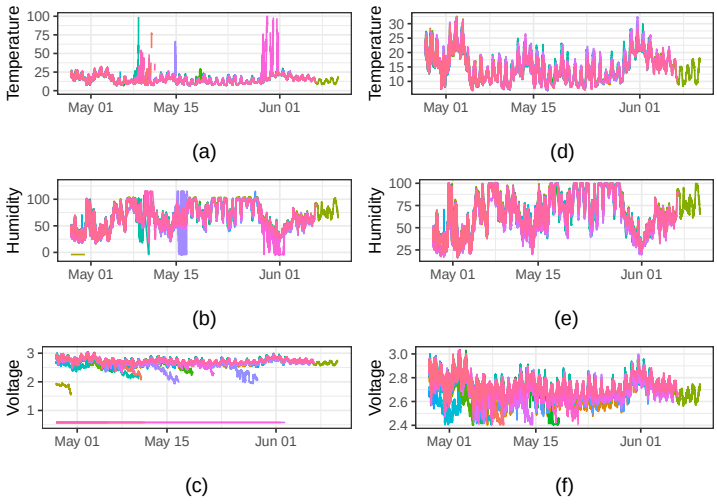

Figure 2: A comparison of removing anomalous sensor readings. Plots (a) (b) and (c) are the readings before removing anomalous data, and plots (d) (e) and (f) are the readings after removing anomalous data.

Since abnormal readings are often results of battery failure, we discovered that filtering out failing voltage also gets rid of outliers in other readings. After converting units to correct range and removing wrong readings, we present the histograms of the four variables of interest in Figure 3. The final data set after clean-up has 262,675 observations from 62 nodes (31 of them recorded in `sonoma_data_net`), whose readings covered a time period from 2004-04-27 17:14:58 to 2004-06-10 13:59:58. No significant outliers exist in the data set, indicating a successful data cleaning process.

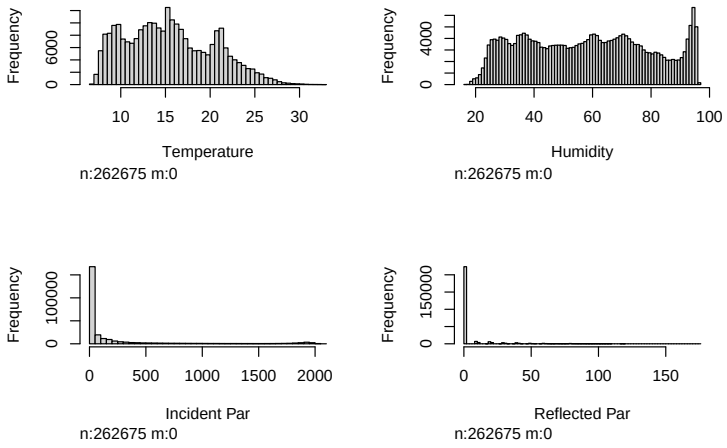

Figure 3: The histograms of four variables of interest (transformation done in Section 5).

3 Data Exploration

Based on Figure 4, available readings diminished towards the end as more and more sensor stopped working. It could be worthwhile to examine the fluctuations of readings throughout one or several days. We decided to choose a time when there were fewer sensor failures and most of the available sensors worked normally. We first selected May 15, 2014 because the numbers of readings collected around that day were consistent and stable.

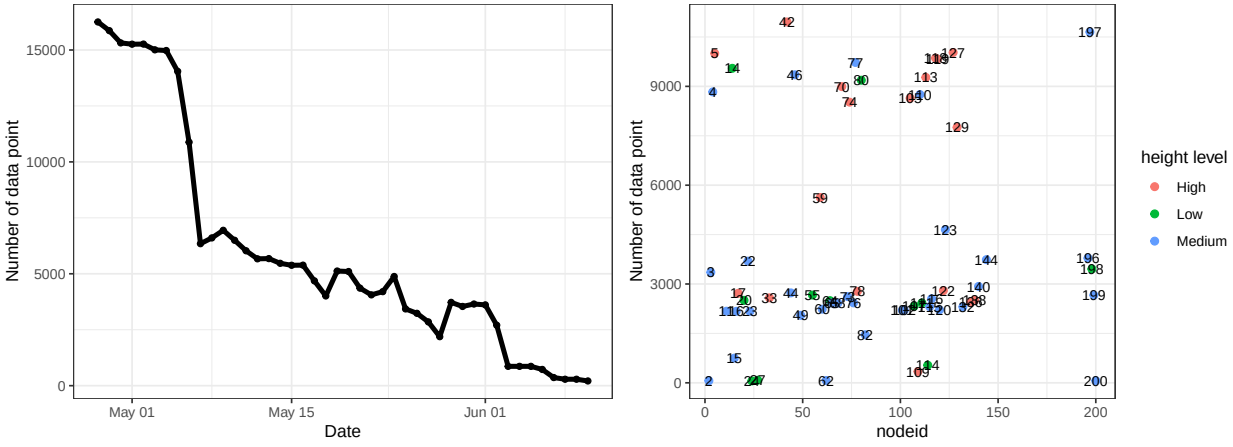

Figure 4: The number of data points collected from each day or each node of the experiment.

Table 1 presents the summary statistics of variables, including their ranges, averages and standard deviations, which are reasonable and coherent with the original paper.

In Figure 5, we found strong correlations between several variables of interest. First of all, a strong negative linear relation occurs between `humidity` and `temperature`. This is intuitive as higher temperature aids water evaporation from ground surface, leading to a reduced humidity. `Incident PAR` and `Reflected`

Table 1: Summary Statistics of Numerical Variables

Statistic	N	Mean	St. Dev.	Min	Max
Temperature	262,675	15.507	4.901	6.778	32.581
Humidity	262,675	58.158	21.378	16.228	96.565
Incident Par	262,675	212.775	469.244	0.000	2,071.426
Reflected Par	262,675	4.888	14.939	0.000	175.570
Height	262,675	47.576	12.455	10.500	66.500

`PAR` have a correlation coefficient of 0.5, a reasonable result since stronger sunlight would increase both readings. Furthermore, `temperature` and `Incident PAR` share a correlation coefficient of 0.5, as exposure to intense sunlight increases both `Incident PAR` and `temperature`. Another interesting observation is the strong positive correlation between `voltage` and `temperature`. This can be explained as higher temperature intensifies chemical reactions within the battery and lead to a higher voltage.

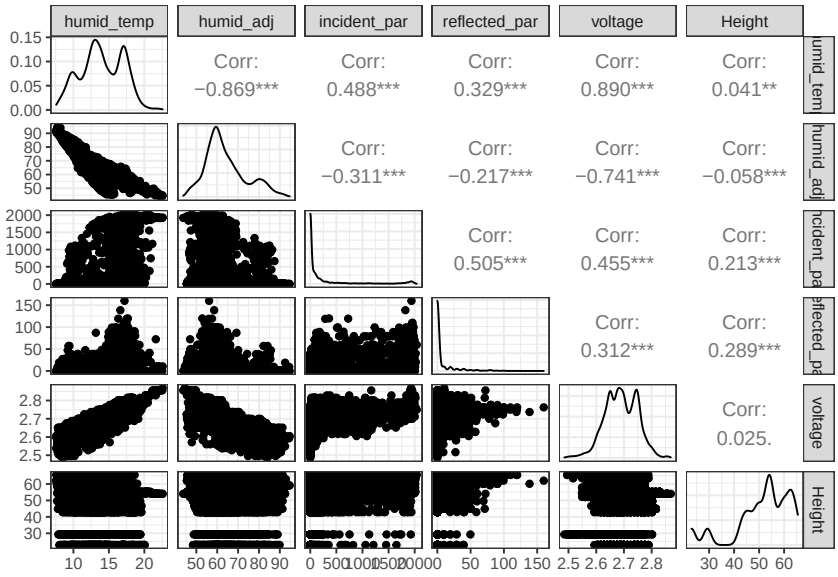

Figure 5: The pairwise scatterplots of variables of interest on May 15.

Before examining the temporal trend, we think `Height` variable could be an important factor influencing node readings. Therefore, we created a new categorical variable indicating `height level`, with 0 to 30 feet being low, 30 to 50 feet being medium and 50 feet and above being high.

We then examine the time series plots of variables of interest within a single day. Figure 5 shows how readings alter with time, with color representing the heights of sensor nodes. Clear time dependency is visible in the plot. After noon, `temperature`, `Incident PAR`, and `Reflected PAR` reach a maximum due to sunlight exposure. `Humidity`, as a result, reaches its daily minimum. After sunset, temperature and PAR values gradually decrease and humidity begins to peak. The effect of `Height` is also observable in the plot. While higher nodes tend to have lower `humidity` readings, `Height` is usually positively associated with higher readings of temperature and PAR values. These plots confirm our successful unit conversion.

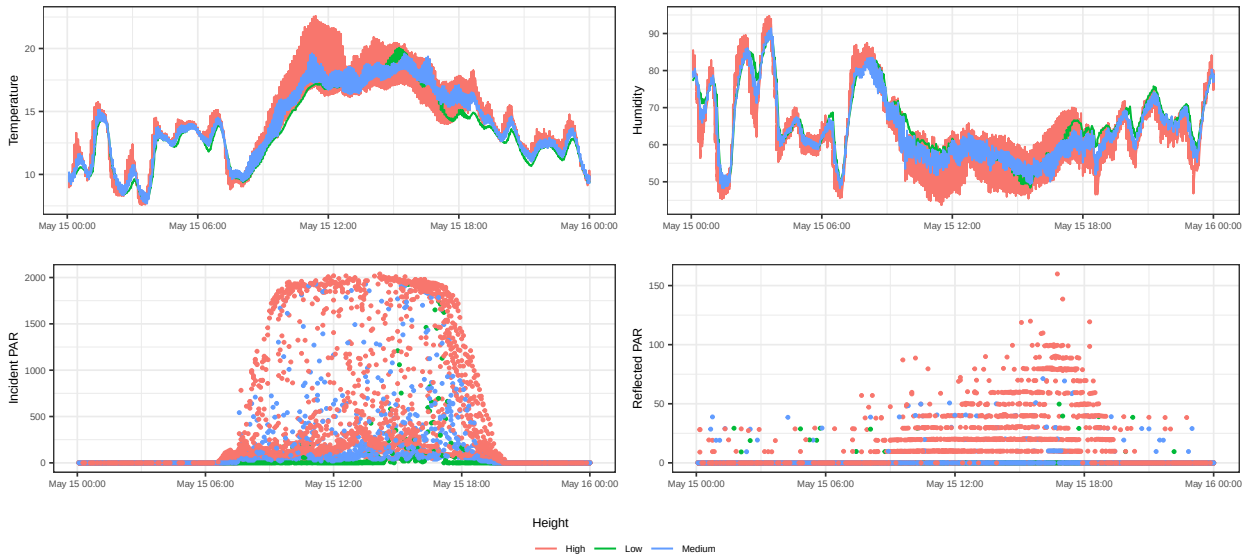

Figure 6: The time series plot of readings on May 15.

We further selected three days with abundant and stable readings, which lasted from April 30 to May 2. Figure 7 shows that these three days experienced a consistent rising of `temperature`, as consequently `humidity` dropped every day. In fact, the highest temperature, 32.58 degrees, during the entire experiment was recorded in the afternoon on May 2. The PAR values were stable across these days since the period was likely to be more and more sunny.

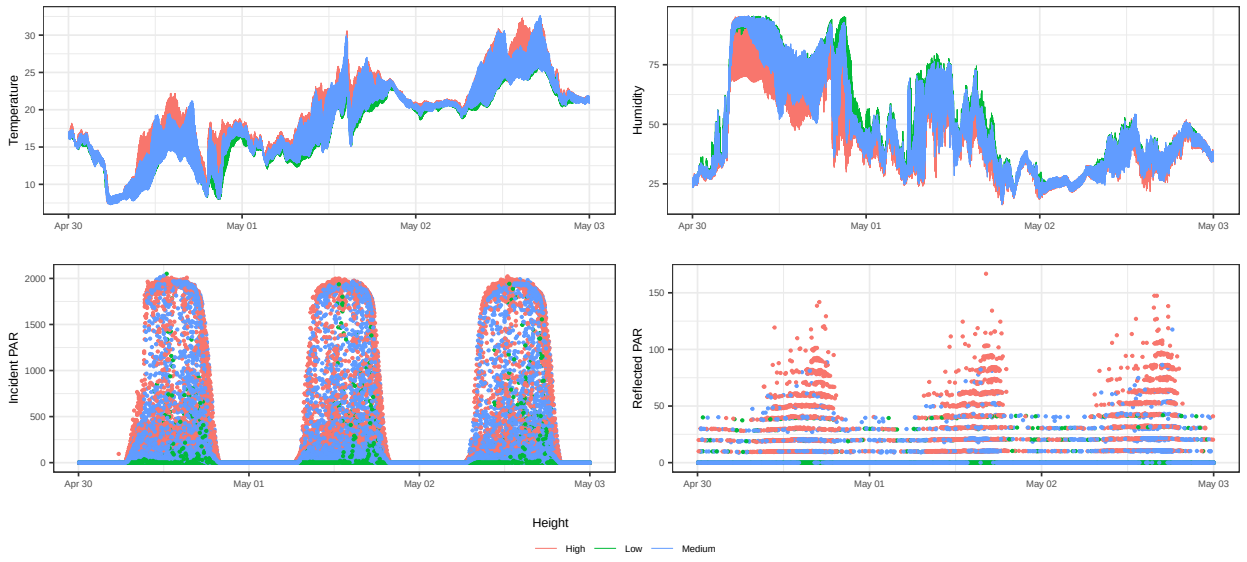

Figure 7: The time series plot of readings from April 30 to May 2.

A Principal Component Analysis (PCA) is conducted on the cleaned data set. From the scree plot below, we can see that two dimensions are enough to explain for 84% of variations within the data. Closely evaluating each PC direction, it is observed that the first PC direction is dominated by `humidity` and `temperature` while the second PC direction is dominated by `Incident PAR` and `Reflected PAR`. This PCA

result matches the pattern discovered in 5. With the high correlation between `humidity` and `temperature` and between both `PAR` values, it is sufficient to reduce the data dimension from four to two.

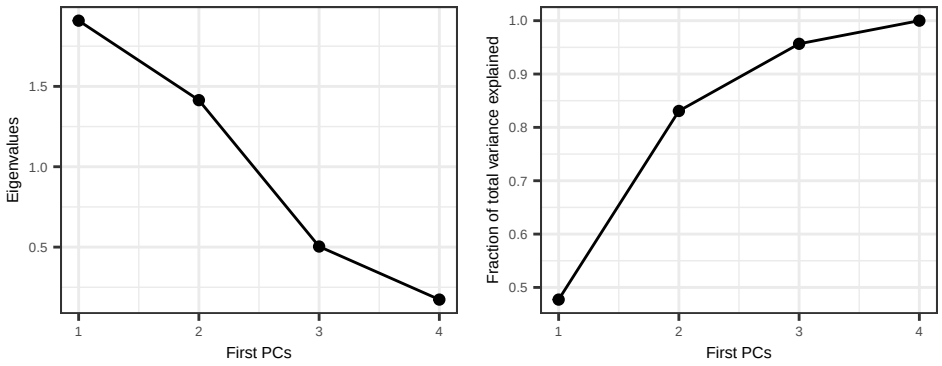

Figure 8: Scree plot and total variation plot suggest that two PCs are sufficient low-dimentional representations to approximate the data with the four key variables of interests.

4 Findings

Detailed findings from our data exploration process is presented in this section. Specific findings include lower dimension representation of data with PCA, day and night data differentiation with clustering methods, clustering of data from different weather condition, and clustering of data from different height.

4.1 Finding 1: Day and Night Data Differentiation with Clustering

With the conclusion of a potential lower dimension data representation from the last section, we formed the hypothesis that data can be clustered with specific clustering algorithms such as EM or K-means. In this part, we try to utilize clustering methods to differentiate data collected from day time and night time. We selected data collected between 0:00 to 1:00 and 12:00 to 13:00 on May 15, 2004, a time range where sensor stability is desirable.

The left plot in Figure 9 displays the original data projected on the two PC directions, with red dots representing night time data and green dots showing data collected during day time. The two plots on the right in Figure 9 display the application of K-means and GMM algorithm to the data points. Both methods are largely accurate with K-means having a marginally better performance. This is resulted from the difference in underlying mechanisms: K-means minimizes the total within-cluster distance while GMM assumes data are generated from some Gaussian distribution. Since the selected data do not perfectly follow a normal distribution, GMM has a worse performance. This abuse of model assumption is further demonstrated in the next subsection. Overall, we illustrate here how the clustering method can be utilized to differentiate data collected during the day time or night time.

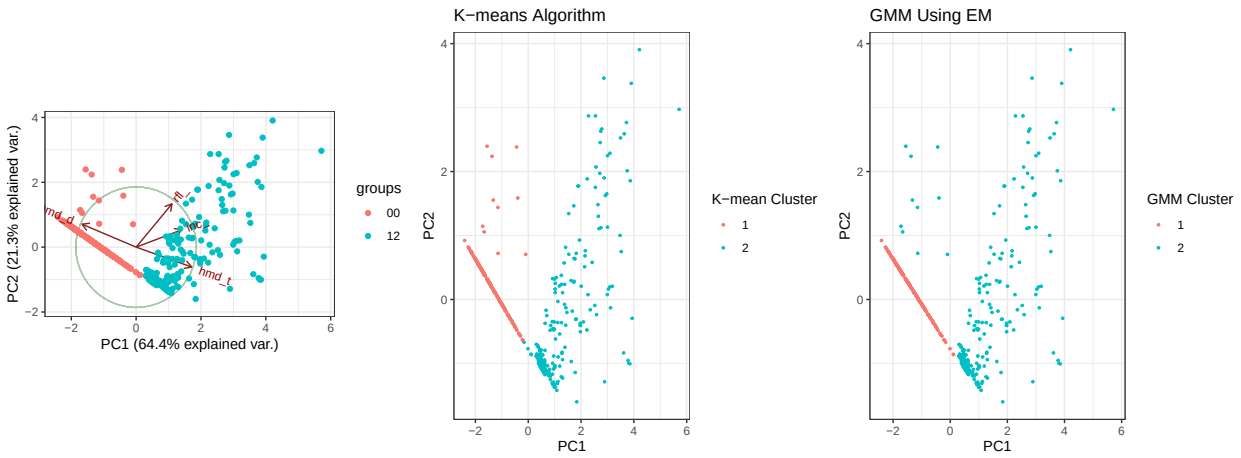

Figure 9: PCA and clustering of data points on May 15, 2004

4.2 Finding 2: Differentiation of Data Collection Date

In this section, we further utilize the two clustering methods to differentiate data collected under different weather conditions (rainy versus sunny). May 2 (sunny) and May 17 (rainy) are selected for analysis. In figure 10, we can observe the clear discrepancy in `temperature` and `humidity` readings between data collected these two days.

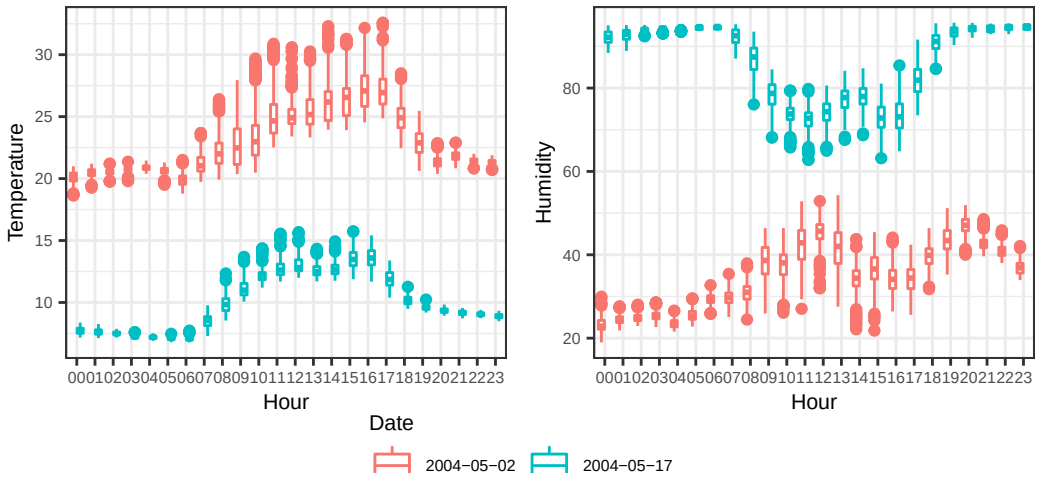

Figure 10: The key difference between May 1 and May 17 lies in their temperature and humidity.

In Figure 11, the projected data is displayed on the left, with May 2rd in red and May 17th in green. A clear clustering pattern can be observed from the PC plot. In the right two plots, K-means and GMM are applied to the data. While K-means generates the correct clustering pattern, GMM model provides a completely wrong result. These two plots demonstrate the consequence of data failing to satisfy Gaussian model assumptions. Taking a closer observation, it can be concluded that the data is generated from a "half Gaussian" mechanism rather than Gaussian. Thus the result of utilizing GMM model is disastrous. K-means, on the other hand, provides a better clustering result. Overall, we demonstrated how K-means can be utilized to differentiate rainy versus sunny days accurately.

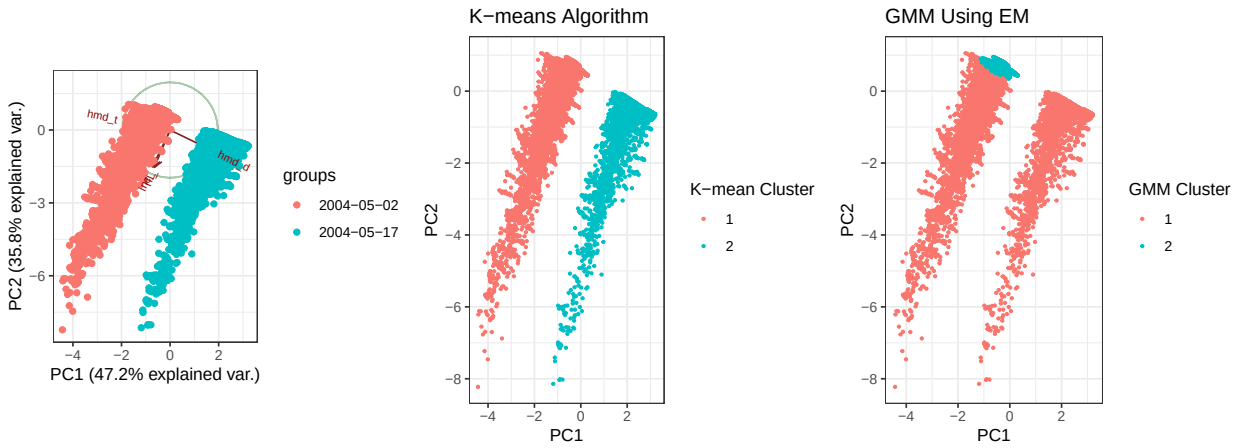

Figure 11: PCA and clustering of data points on May 1 and May 17

4.3 Finding 3: Differentiation of Data from Different Heights

In the last two sections we discovered that K-means algorithm outperforms GMM when applied to clustering of data collected under different time and weather condition. In this section, we explore the circumstance where GMM can be utilized to distinguish the height levels of sensor locations. Data collected around noon by node 14 (at 29 feet height) and node 42 (at 59 feet height) is selected. We want to investigate if `Height` has a significant effect on sensor readings. Data is projected onto 2 PC directions in Figure 12, with red representing node 14 and green displaying node 42. Application of K-means and GMM are displayed in the right two plots. K-means presents a totally wrong clustering pattern while GMM is able to correctly differentiate data, a result of underlying Gaussian data generation mechanism. From the clear two clusters, we can conclude `Height` has a determining effect on environmental data surrounding the tree, especially on Incident and Reflected PAR. Figure 14 displays the positive linear relation between `Height` and `Temperature`, `incident par`. A negative linear relation exists between `Height` and `Humidity`.

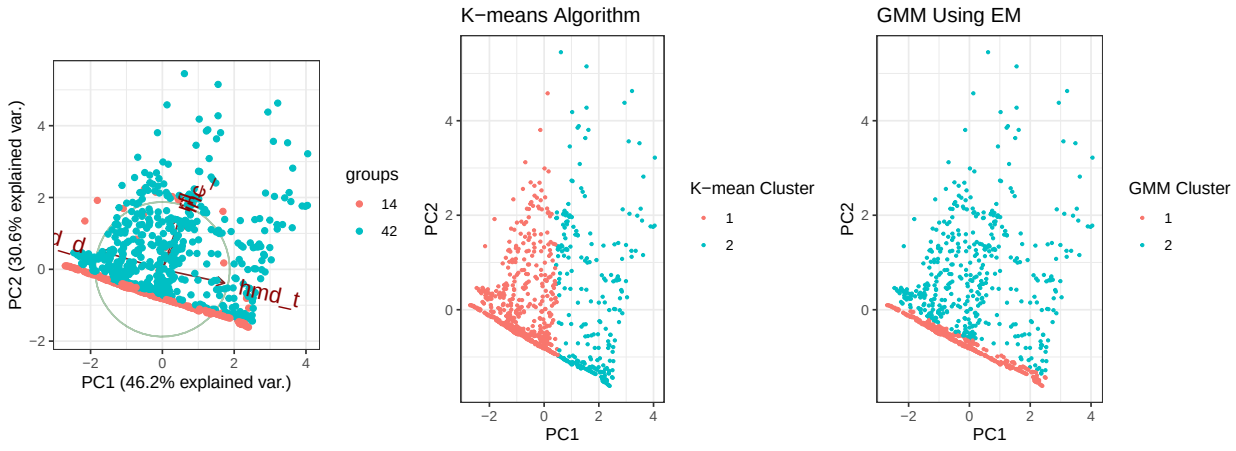

Figure 12: PCA and clustering of data points from node 14 (at 29 feet high) and 42 (at 59 feet high)

5 Graph Critiques

Despite the sound quality of the original paper, we found some potential improvements for data visualization, including log transformation of data, changing data range, altering data axis, etc.

The first potential improvement is a log transformation of histogram data in Figure 3[a]. Since both incident and reflected PAR has a long tail, data are heavily clustered towards one end. Useful information about data distribution can not be easily accessed from the original histogram. As an alternative, a histogram with log-transformed data is plotted below. It can be observed that the long tail is no longer present and the audience can have a better sense of data distribution.

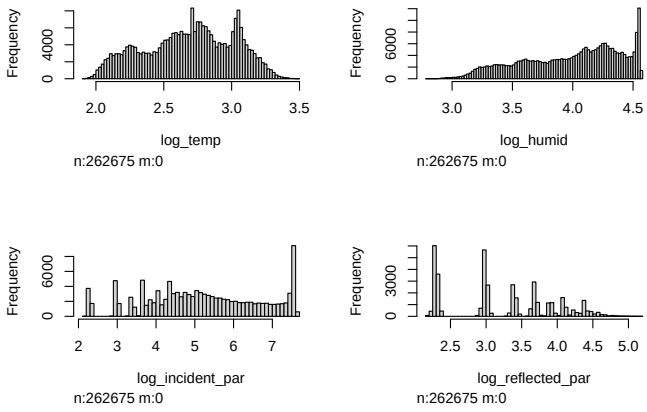

Figure 13: Histograms for variables of interests with log transformation.

Another potential improvement of graphical visualizations are over Figure 3[c] and Figure 3[d], where the author tries to present the distribution of sensor reading at each height. However, since `Height` is an independent variable here, it should be on the horizontal axis instead of the vertical axis. The original plots are misleading as it take height as the dependent variable. We also learned from Figure 4 of the original paper to select just one day's data instead of plotting all days. A better version would be:

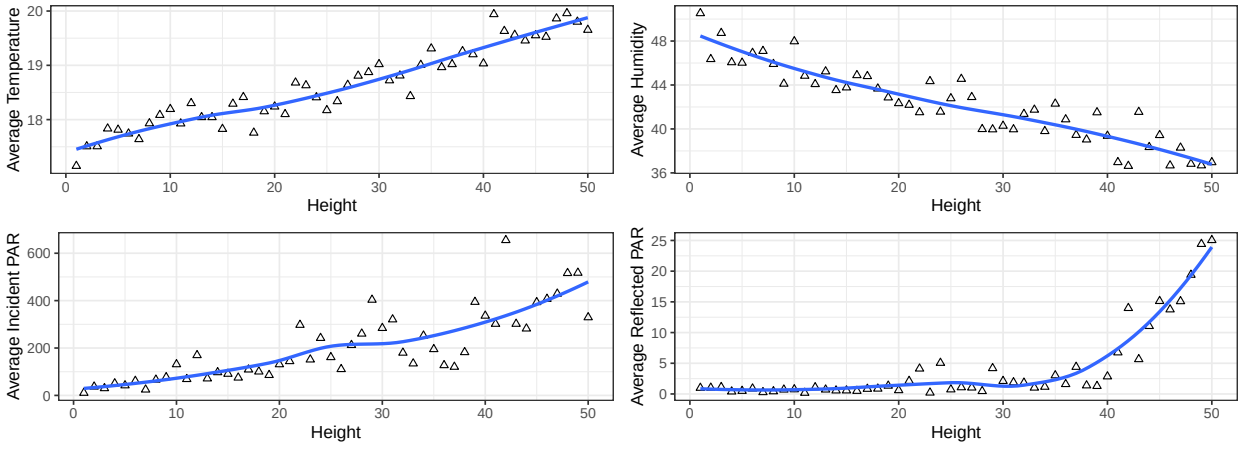

Figure 14: Adjusted visualization between Height and other variables on May 1.

Furthermore, the first two plots of figure 4 can also be enhanced. While the author utilizes different colors to represent readings from different nodes, these colors are indistinguishable. Readers have no information about which color represents which node or the specific height of data collection. Here coloring becomes merely a visual effect. An improvement could be to compute the average readings at each time point, and turn `Height` into a factor variable as described in Figure 6 and use it for coloring instead. Through this transformation, we can add one more dimension to the plot and provide readers with more information.

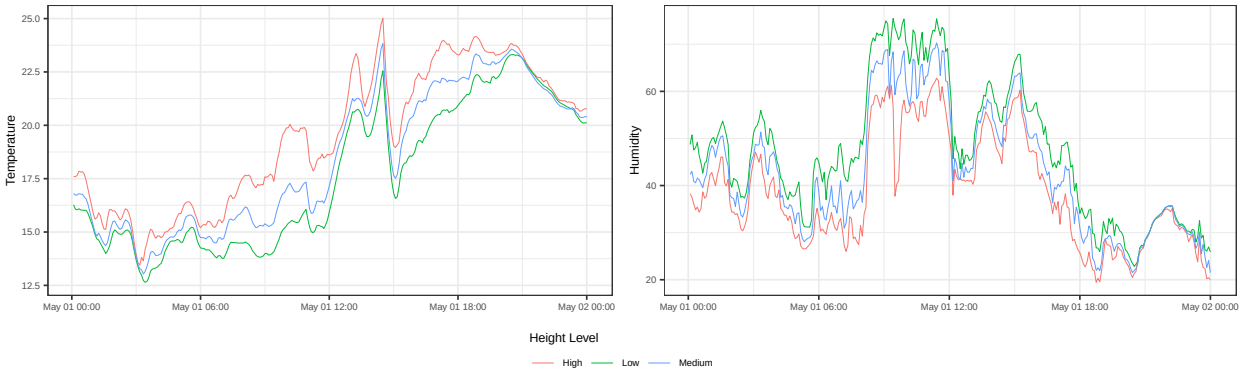

Figure 15: The time series plot of readings on May 1.

Last but not least, modifications can be made to Figure 7 to enhance informativity. Even though the author plotted the percentage yield, no information is provided concerning total yield amount across different heights and dates. Such information is beneficial for readers to understand detailed data collection process. Furthermore, side-by-side comparison plots can be employed to highlight the difference between net and log data. To improve these two points, we propose using the following plots:

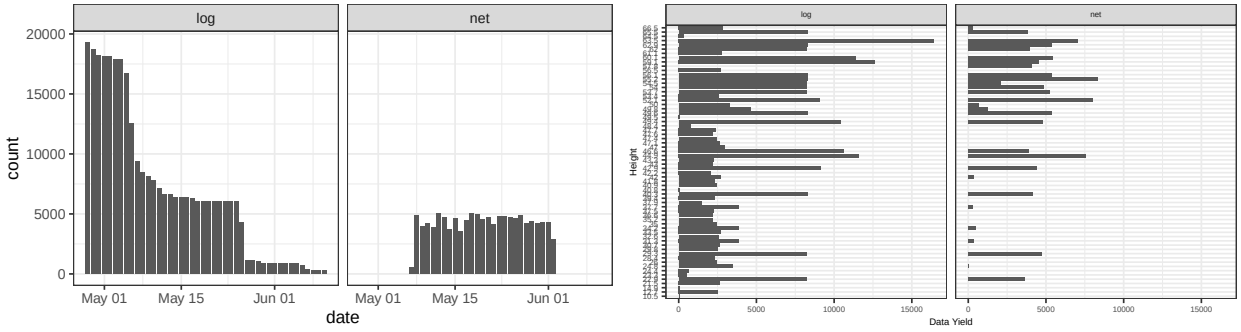

Figure 16: The amount of data yielded by log and net data.

6 Reference

Tolle, Gilman, Joseph Polastre, Robert Szewczyk, David Culler, Neil Turner, Kevin Tu, Stephen Burgess, et al. 2005. "A Macroscope in the Redwoods." In *In Proceedings of the 3rd ACM Conference on Embedded Networked Sensor Systems (SenSys)*, 51–63. ACM Press.