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 position of sensor 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 interests, and finally removing apparent outliers and calibrating values from sensors. They concluded their work by elaborating 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 bilogical 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 the whole 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 the 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 identity the battery failure time as depicted in the paper. However, utilizing the voltage unit from `sonoma-data-net.csv`, we see a 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 reading 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 matches readings in the log data file. Similarly significant, each value have a lower frequency in the net file than the log file, indicating the conversion is successful.

A 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

Another crucial element of the data cleaning process is the restoration of time variable. In the file `sonoma-data-log.csv`, the time variable is fixed at 2004-11-10 14:25:00 which is the data extraction time. For data analysis purpose, we want to restore the time variable to actual data collection time. This is accomplished by evaluating the `nodeid` variable and `epoch` variable. 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 the nighttime. We thus crossed-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.

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 guidence 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 interests, 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 from 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.

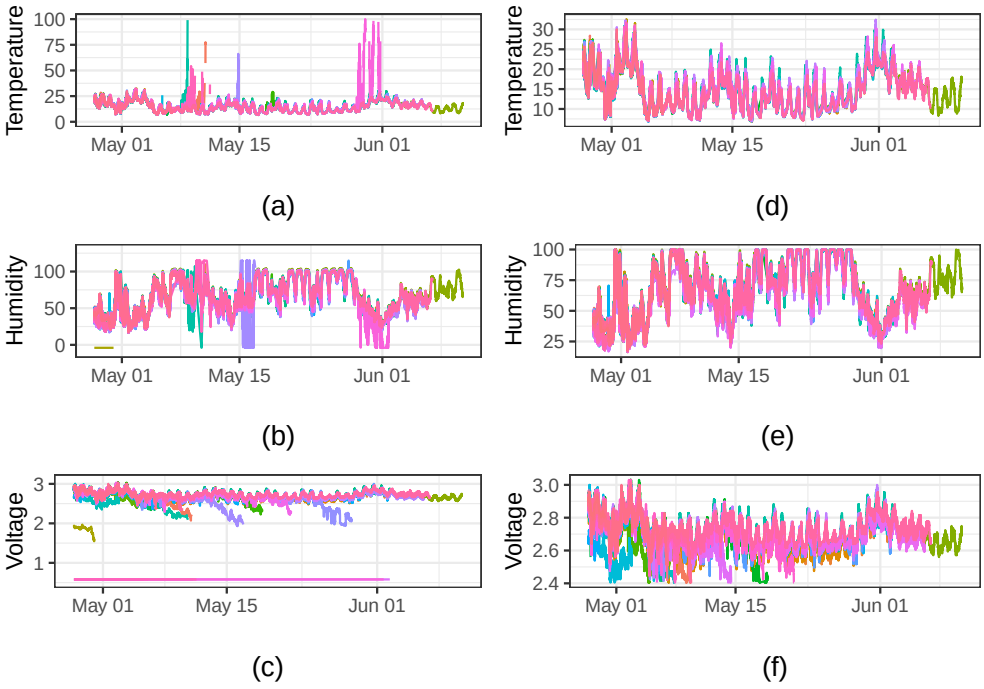

Figure 2: A comparison of removing anomalous sensor readings. Plots (a) (b) and (c) are the readings before removing anomalous data, and plots (d) (e) and (f) are the readings after removing anomalous data.

After converting units to correct range and removing wrong readings, we present the histograms of the four variables of interests in Figure 3. The final data set after clean-up has 262,675 observations from 62 nodes, whose readings covered a time period from 2004-04-27 17:14:58 EDT to 2004-06-10 13:59:58 EDT.

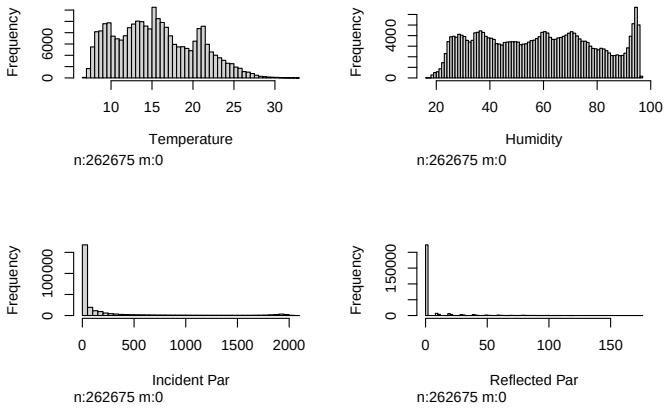

Figure 3: The histgrams of four variables of interets

3 Data Exploration

It could be worthwhile to examine the fluctuations of readings throughout a single day. We decided to choose a time when there were few sensor failures and most of the available sensors worked normally. Based on Figure 4, we selected May 15, 2014 because the numbers of readings collected around that day were consistent and stable.

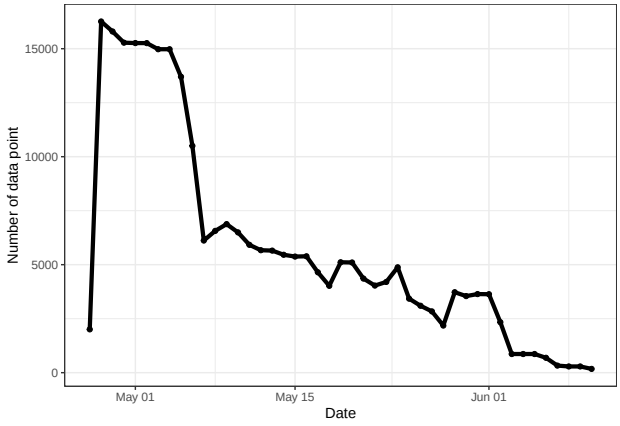

Figure 4: The number of data points collected by each day of the experiment

From Figure 5, we found strong correlations between *need to add more content here*

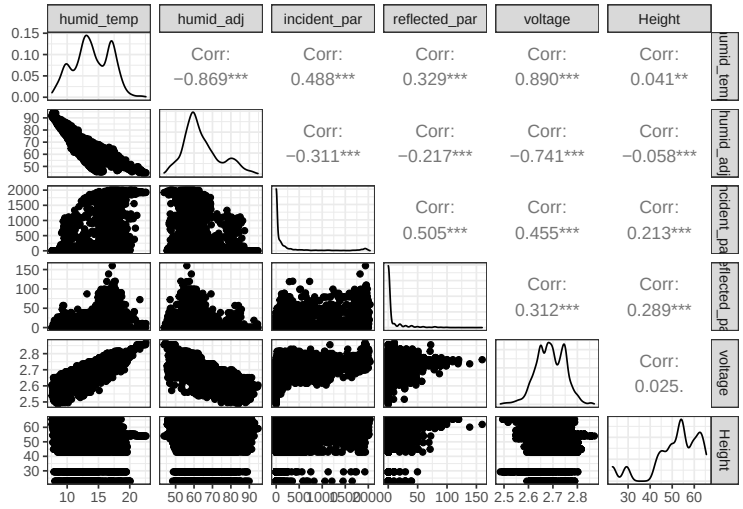

Figure 5: The pairwise scatterplots of variables of interests

We also examined that time series plots of variables of interests within a single day. Figure 5 shows that *need to add more content here*

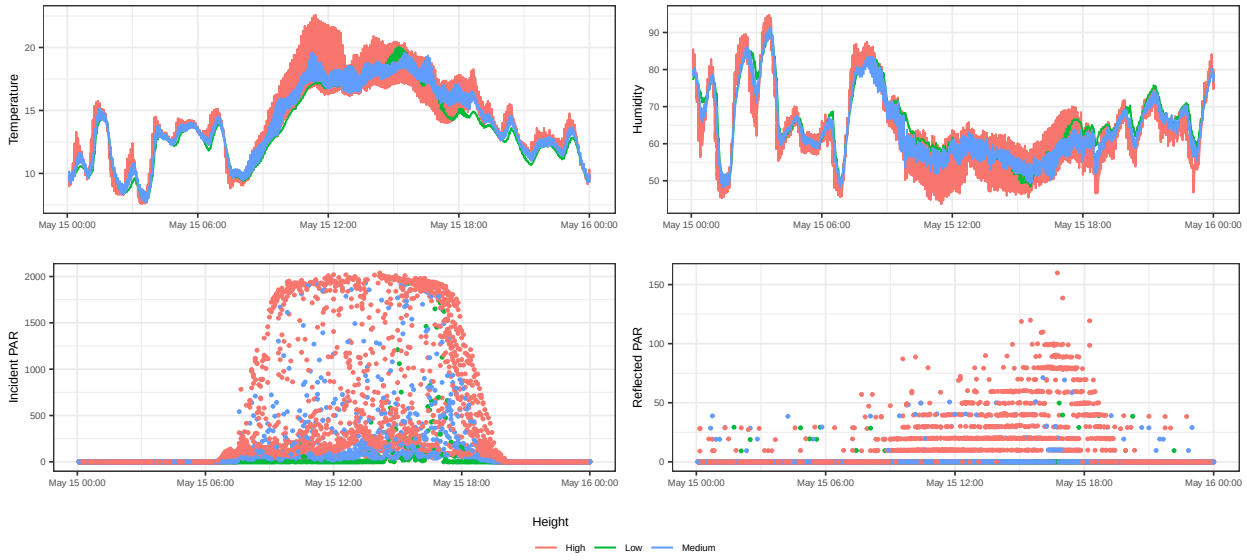

Figure 6: The time series plot of one day's readings

Do we want to use 3 day? (Plots are ready here) Figure 7

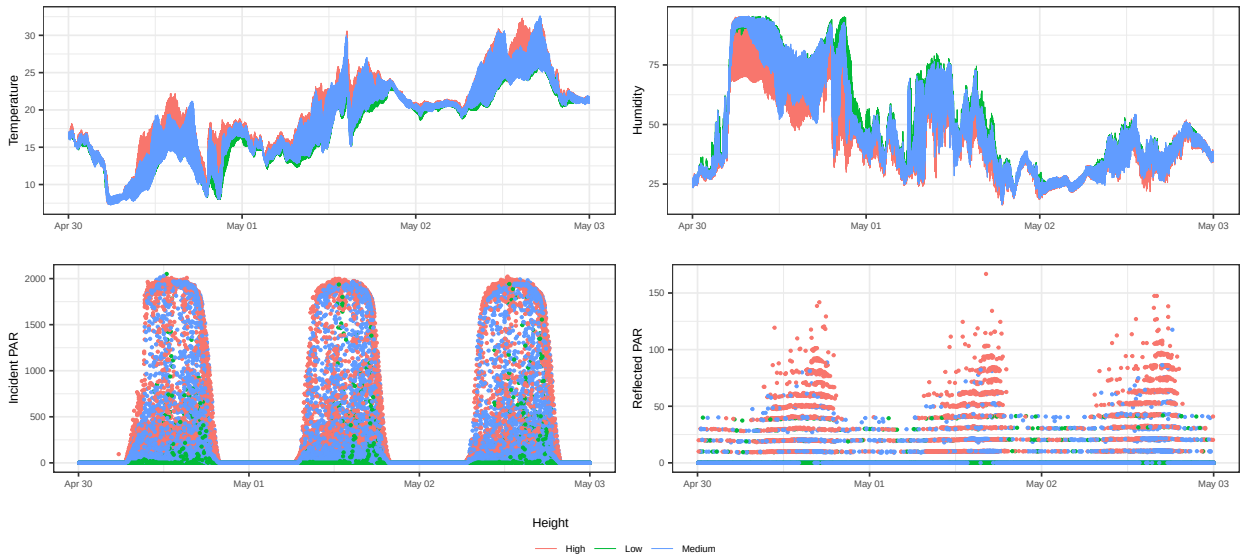

Figure 7: The time series plot of three days' readings

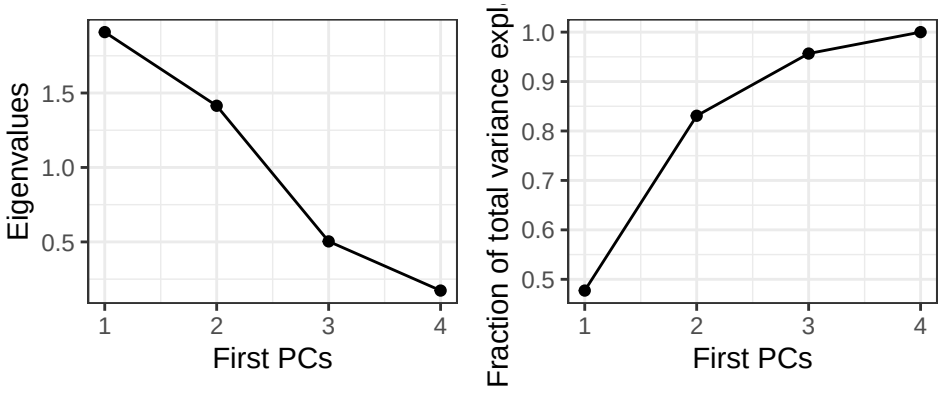

Figure 8: Scree plot and total variation plot suggest that two PCs are sufficient low-dimentional representation to approximate the data with the four key variables of interests

4 Findings

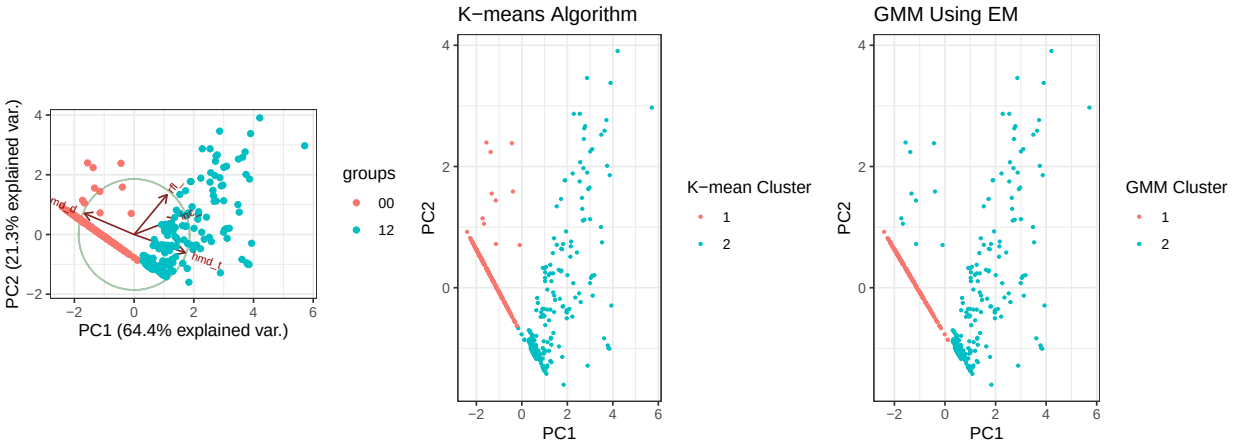

Figure 9: PCA and clustering of data points on May 15, 2004

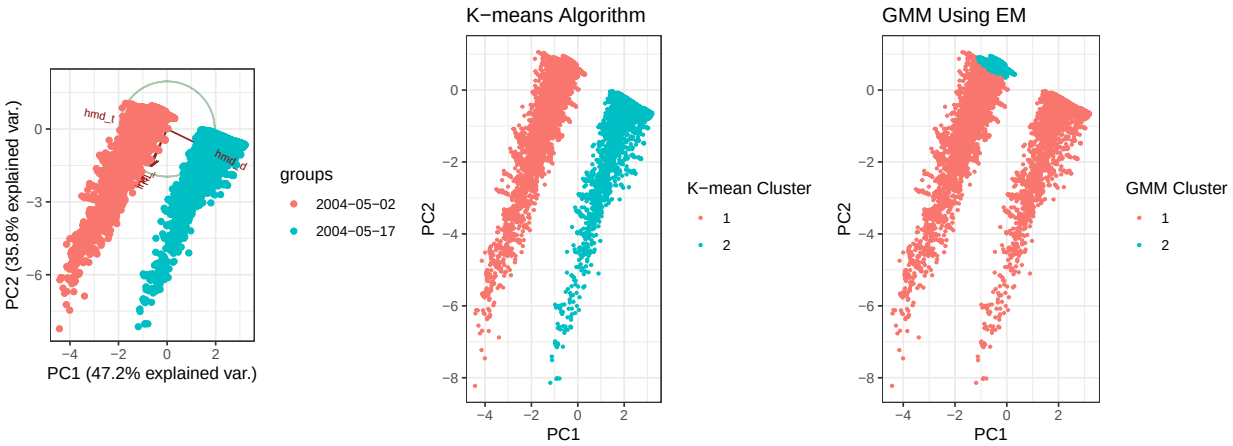

Figure 10: PCA and clustering of data points on May 1 and May 17

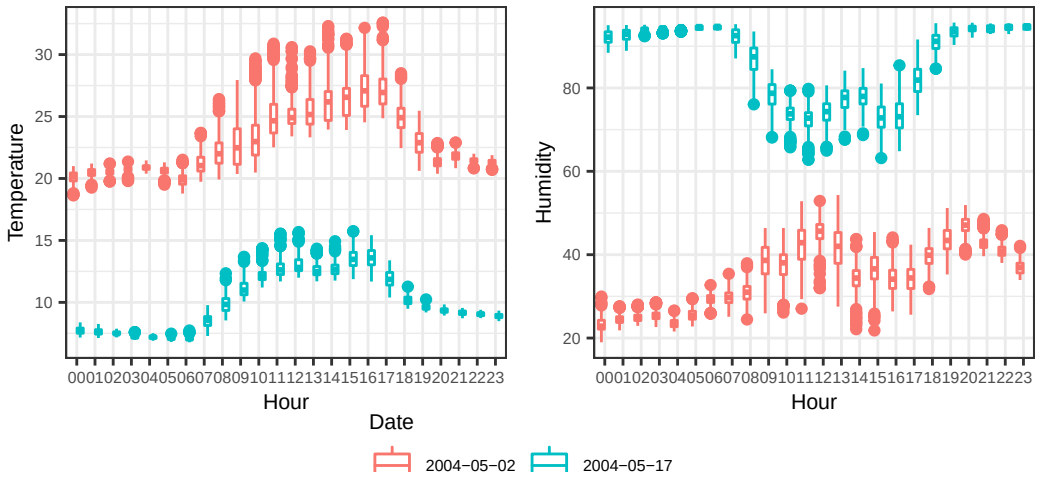

Figure 11: The key difference between May 1 and May 17 lies in their temperature and humidity.

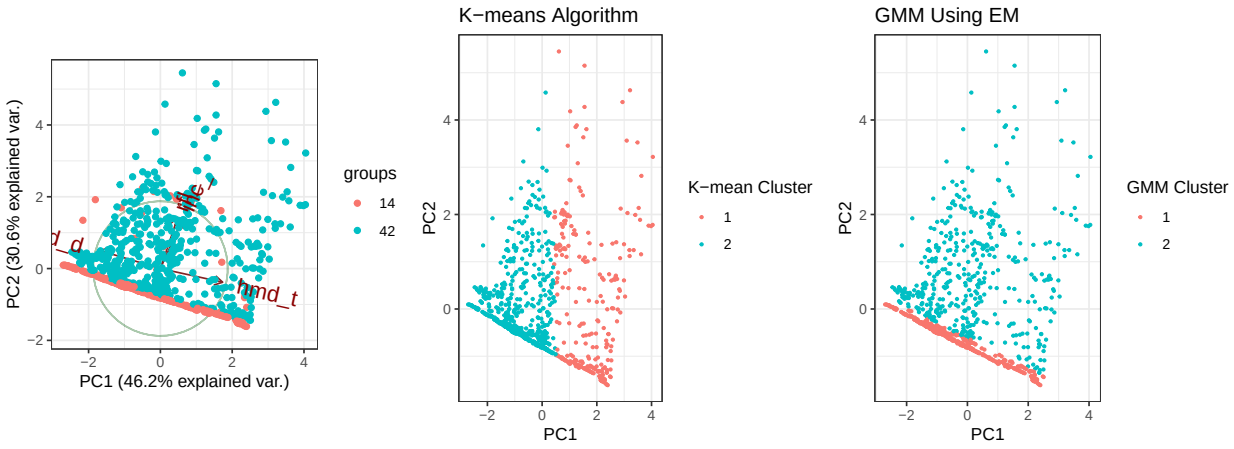

Figure 12: PCA and clustering of data points from node 14 (at 29 feet high) and 42 (at 59 feet high)

5 Critiques

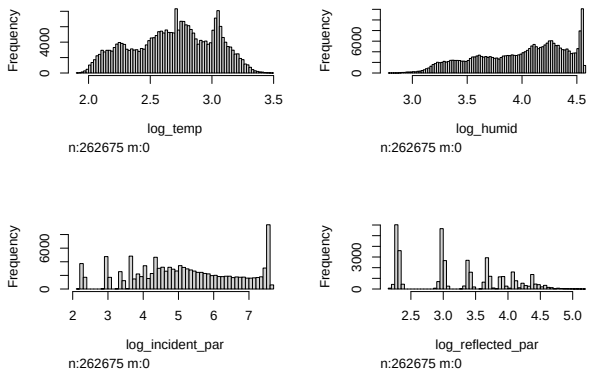

Figure 13: Histograms for variables of interests with log transformation.

6 Appendix

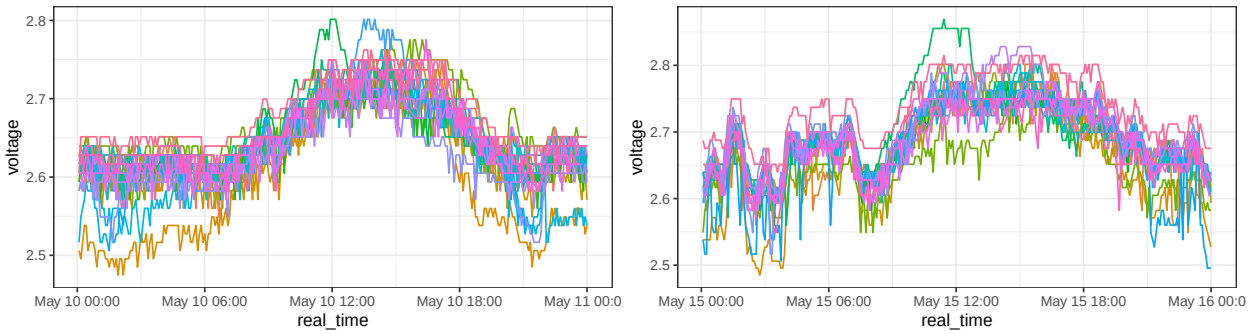

Figure 14: Daily voltage readings on May 15 and May 20.

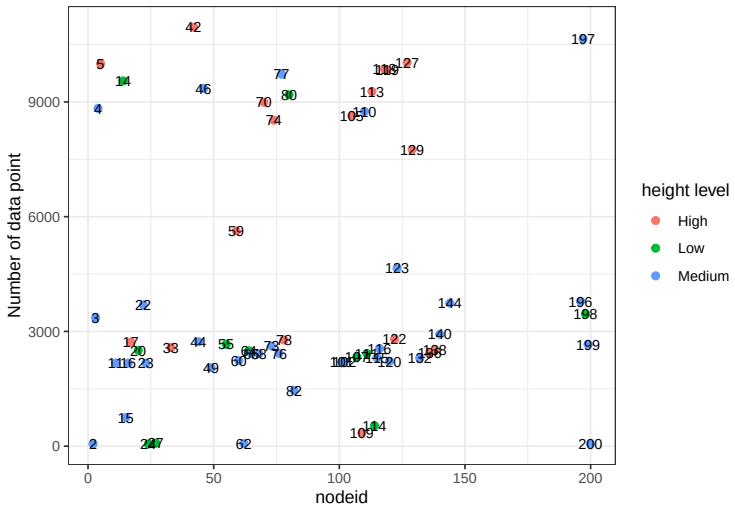

Figure 15: The number of data points collected by each functional node.

7 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.