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

We noticed the units of `voltage` are not consistent between `sonoma-data-net.csv` and `sonoma-data-log.csv`, with the first requiring analog-to-digital conversion (ADC) and the latter using normal voltage units. The `voltage` conversion formula:

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

We also converted the units of `hamatop` and `hamabot`. The sensors recorded the two PAR measurements in Lux, and we converted to the unit of PPFD (μmol $m^{-2}s^{-1}$) by a conversion factor of 54 because the light Source is assumed to be mostly sunlight.

2.2 Outlier Rejection

In terms of outliers, 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 1 presents the comparison of before and after removing anomalous sensor readings.

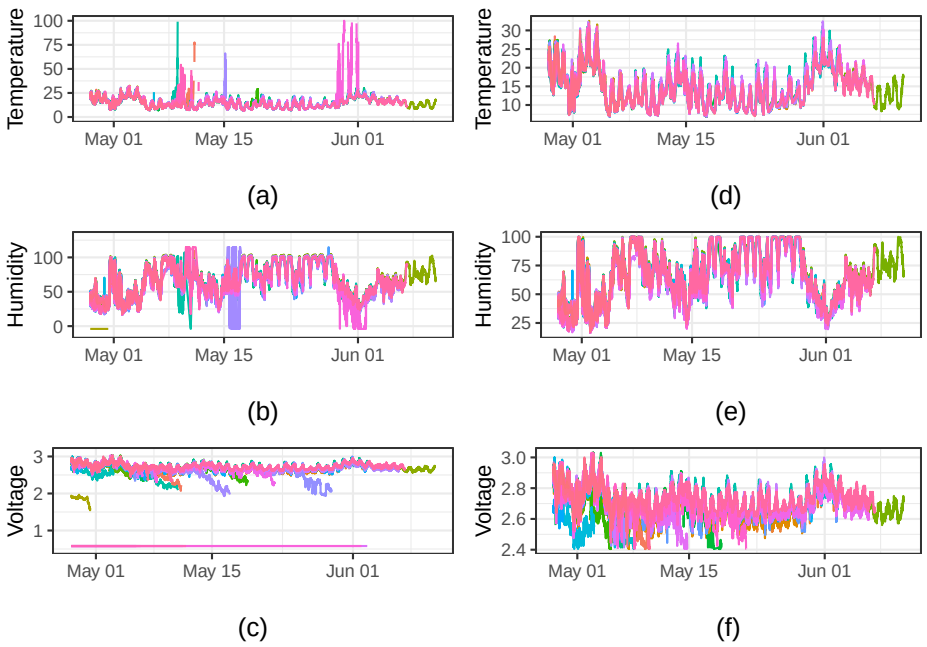

Figure 1: A comparison of removing anomalous sensor readings. Plots (a) (b) and (c) are the readings before removing anomalous data, and plots (d) (e) and (f) are the readings after removing anomalous data.

3 Data Exploration

4 Findings

5 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.