Project 1 Redwood Data Report

Phoebe Abramowitz(26386343)
Omri Newman(3032273024)

Due by: **11 PM, Friday, Mar 8**

1 Paper Summary (20 pts)

This study of microclimatic monitoring aims to get a better understanding of the life of a Redwood tree in Sonoma, CA. To do this, an interdisciplinary team from the University of California, Berkeley designed an experiment that paints a picture of a redwood tree's ecophysiology spanning fourty-four days in Spring 2004. They did this using a unique network of nodes and wireless sensors to gather spatio-temporal and environmental dynamic data over the entire organism. Each of the thirty-three nodes collected hundreds of thousands of data points, which lie in four different three-dimensional spaces (time x height x value). This wireless sensor network, or macroscope, allows for rich analysis on an unprecedented dataset. After brainstorming with several biologists, Sonoma's coastal redwood forest was chosen as the site of this study because of its dynamic ecosystem. Longer daylight hours during the Spring provided sufficient time for the photosynthetically active radiation sensors to collect data, and the coastal fog coming off the Pacific guaranteed a wide range of temperature and humidity gradients. A thorough analysis of the redwood data confirmed the biologists' hypothesis, which premised the existence of dynamic gradients present in the redwood trees ecophysiology. By projecting the data points onto a subset of the three-dimensions, the analysis task became simpler. One drawback to using the sensor network was the presence of missing values throughout the data set in both time and height dimensions. However, the team worked in spite of this by studying distributions in most of their analyses. In addition to confirming the team's hypothesis, more information about the climatic distribution of the tree was extracted from the analysis. One impactful step was using the data to build a quantitative model of the effect of microclimatic gradients on the sap flow rate of redwood trees. By better understanding this process, biologists can piece together a more detailed picture of the larger scale carbon and water exchange within a forest ecosystem. The main variables of interest in this study are those that shed light on the microclimate and ecophysiology of the redwood tree. Thus, each data point can be viewed in three dimensions– time, height, and value. The values, recorded by the network of sensors, are temperature, humidity, incident photosynthetically active solar radiation (PAR), and reflected PAR. Collecting this data over an extended period of time required a powerful system and precise deployment methodology. The data collection time period started in late April and spanned fourty-four days, sampling all sensors every five minutes. The nodes were placed along the seventy meter redwood tree starting at a height of fifteen meters, with two-meter spacing between nodes. The west side of the tree had a thicker canopy that served as protection against environmental effects, so the majority of the nodes were placed there. Lastly, the nodes were placed close to the trunk (0.1-1.0 meter) such that the recordings were of the microclimatic trends of the tree and not the wider climate. Multiple sensors were integrated into the Mica2Dot wireless node, as a means to gather data on the environmental dynamics of the tree. This node has a one inch diameter form factor, an Atmel microcontroller running at 4 MHz, a 433 MHz radio from Chipcon operating at 40Kbps, and 512kB of flash memory. The node was packaged into a sealed cylindrical enclosure made from white high-density polyethylene that reflects radiated heat. The enclosure contained the node, battery, and two sensor boards - one for direct radiation and one for all other measurements. Data collected from these nodes were stored in a local database or gateway, then transmitted to an offsite database via a general packet radio service (GPRS) cellular modem. The UCB team used TASK software as the node operating system and data collection scheme to aid with this work. This multi-hop system utilized TinyOS and MintRoute systems to retrieve data from the nodes and store them in the gateway as efficiently as possible. The entire network of nodes was awake for four seconds every five minutes, in order to take sensor readings, transfer the data to the base stations, and then return to sleep. Two calibration schemes were completed by the team to ensure a satisfactory operation prior to setting up the final system. The first calibration consisted of leaving the sensors on a roof collecting PAR data every thirty second for two weeks to verify the accuracy of the readings. The purpose of the second calibration was to understand the response of the temperature and humidity sensors. The nodes were tested in a controllable weather chamber that had varying temperature and humidity conditions, and recorded readings every thirty seconds. The calibration techniques used on the nodes took up space on the 512kB flash memory. Thus, some of the data loggers in the local database gateway ran out of memory before the end of the data collection time period. These loggers were meant to serve as a backup in case of network failure, and stored data in the case of lost GPRS connectivity. On the other hand, the logger data that was stored in the gateway could only be accessed at the end of the data collection time period, and there was lost information in the nodes that ran out of flash memory. This defines the main difference between sonoma-data-net.csv and sonoma-data-log.csv, which both contain values the other is missing.

2 Data cleaning (40 pts)

This data set contains gross outliers, inconsistencies, and lots of missing values. We use basic exploratory data analysis and data cleaning techniques to get a better understanding of the data, determine which values make sense in the context of the study, and remove missing values when necessary.

(a) For the initial comparison in hamatop (Incident PAR) between **sonoma-data-log.csv** and **sonoma-data-net.csv**, we remove outliers that are ten times larger than the third quartile in an effort to explore the ranges more thoroughly. We notice a slight bump in the right tail of incident PARs histogram for **sonoma-data-net.csv**, which does not occur in **sonoma-data-log.csv**. The histograms for hamabot (Reflected PAR) reveal some interesting trends that both **sonoma-data-log.csv** and **sonoma-data-net.csv** share. Not only are the ranges the same between the two, but they both exhibit the same gaps in the data. It is clear from the histograms below that voltage readings are not consistent between **sonoma-data-log.csv** and **sonoma-data-net.csv**.

To reconcile the inconsistencies between voltage and PAR, we applied two separate conversions as to match the ranges from the paper:
The voltage readings in **sonoma-data-log.csv** are all within the correct range of just a few volts, while the readings in **sonoma-data-net.csv** had significantly higher values. This discrepancy can be attributed to the analog-to-digital converter used within the ATmega128 node [1]. To switch from the 10-bit ADC readings to actual voltage measured, we followed a simple ratiometric conversion scheme [2]. We multiplied the values by 12.33 and divided by 1023 which transformed the data into the desired range.

We also noticed the ranges for hamatop and hamabot differ from the paper. We convert the measurements of both types of PAR from Lux to PPMG, the units expressed in the visualizations in the study which requires a 0.0185 conversion factor.

While they are not shown here, the histograms of temperature and humidity revealed no inconsistencies in the ranges, only some outliers.

(b) Remove missing data. Comment on the number of missing measurements and the corresponding date and time period.

In our main data table, we remove 3686 rows where the temperature, humidity, reflective PAR, and incident PAR are missing. There are 3686 values missing for each of the 4 variables.

```
> #Number of missing values for each variable(don't display this)
> sum(is.na(df$temp))

[1] 4262

> sum(is.na(df$humid))

[1] 4262

> sum(is.na(df$incident_PAR))

[1] 4262

> sum(is.na(df$reflect_PAR))

[1] 4262

> #Initial Number of rows missing any measurements
> na_vals <- df %>% filter(is.na(temp)) %>% filter(is.na(humid)) %>%
+   filter(is.na(incident_PAR)) %>% filter(is.na(reflect_PAR))
> #we dont want it to evaluate this value
> nrow(na_vals)

[1] 4262

> #Remove missing Data rows
> df <- df %>% filter(!is.na(temp)) %>% filter(!is.na(humid)) %>%
+   filter(!is.na(incident_PAR)) %>% filter(!is.na(reflect_PAR))
```

(c) After removing missing data, reconciling inconsistencies, and joining the **sonoma-data-log.csv** and **sonoma-data-net.csv**, we incorporate the **mote-location-data.txt** using nodeid as a unique identifier. All together we have the location of each mote within the two trees, along with the respective sensor data. We selected only the variables of interest amongst the 11 to create a desired dataset of all the readings. Our main table has 337,744 observations and 12 variables.

(d) Use histogram and quantiles to visually identify easy outliers for each of the four variables: humidity, humid temp, hamatop, hamabot. And remove them. Comment on the rationality behind your removal.

By noting the ranges of each variable in the paper, we...

(e) (Bonus) Discuss other possible outliers and explain your reason why it is better to remove them than to keep them.

3 Data Exploration (40 pts)

(a) Make some pairwise scatterplots of some variables. Pick a reasonable time period. Explain your choice and describe your findings.

(b) Are any of the predictors associated with Incident PAR? If so, explain the relationship.

(c) Each variable of our data basically have three dimensions: value, height and time. Consider each variable as a time series and look at its temporal trend. Generate such plots (value vs time) with height as color cue for at least four variables (Temperature, Relative Humidity, Incident PAR and Reflected PAR). You can do it for different time scales (during an hour, during a day or during the entire experiment). However, at least the plots with days as x-axis are required. Comment on the range, continuity and strange behaviors in these variables.

(d) After PCA analysis, generate scree plot of the data. Can this data be approximated by some low-dimensional representation?

4 Interesting Findings (15 * 2 pts)

Describe two/three interesting findings from exploratory analysis of the data. Try to use the techniques that you have learned, such as histograms, PCA, K-means, GMM and hierachical clustering etc. Comment on your interesting findings. Different bonuses are given based on how interesing your result is.

(a) Finding 1.

(b) Finding 2.

(c) (Bonus) Finding 3. Bonus is given only if we also find it interesting.

5 Graph Critique in the paper (40 pts)

The overall quality of the paper by Tolle et al. is good. However, some plots are not perfect from a statistician's point of view.

(a) Figure 3[a] shows the distributions of sensor readings projected onto the value dimension, using a histogram. It turns out that both the incident and reflected PAR have long tail. We could not read full information from this histogram. Try to make a better plot with log transform of the data.

(b) What message do the boxplots in Figure 3[c] and 3[d] try to convey? Do you think the plots convey the right messages? If not generate a new plot with the same data. Hint: compare to some plots in Figure 4.

(c) Any suggestions for improving the first two plots in Figure 4? Can you distinguish all the colors in these two plots?

(d) Comment on Figure 7. Is it possible to generate a better visualization to highlight the difference between network and log data?

1: ”10-bit Atmel Microcontroller with 128KBytes In-System Programmable Flash” http://ww1.microchip.com/downloads/en/DeviceDoc/doc2467.pdf
2: ”Analog to Digital Conversion” https://learn.sparkfun.com/tutorials/analog-to-digital-conversion/relating-adc-value-to-voltage