Lab 1 - Redwood Data, Stat 215A, Fall 2017

September 14, 2017

1 Introduction

Wireless sensor networks enable us to monitor the world at a macroscopic level by providing temporal and spatial measurements of various parameters of interest of the study subject. In this report, I will analyze the microclimatic monitoring data of a rewood tree obtained by such wireless sensor networks. The redwood study measured temperature, relative humidity, and photosynthetically active solar radiation of a 70-meter tall redwood tree over 44 days [1]. The dense set of measurements over both space (i.e. length of the tree) and time provides us an opportunity to gain insights into the environmental dynamics surrounding the tree. In this report, I describe the data cleaning process and explore some of the trends revealed by the data, which can potentially be pursued for further study.

2 The Data

2.1 Data Collection

Eighty sensors were deployed on a 70-meter tall redwood tree in Sonoma, California, to measure the parameters relevant to the microclimate around the tree. The measured parameters include temperature (Celsius), relative humidity (percent), incident photo synthetically active radiation (PAR), and indirect PAR. PAR measurements provide information about how much energy is available for photosynthesis. The data was collected over 44 days, from April 27th 2004 at 5:10pm to June 10th 2004 at 2:00 pm, with measurements recorded every 5 minutes. The lowest sensor was deployed at 15m above ground and the highest sensor at 70m, with about 2-meter spacing between the sensors. The majority of the sensors faced the west side of the tree and were deployed 0.1-1.0m from trunk. Several sensors were placed outside this range in order to monitor the microclimate in the immediate vicinity of the tree.

In addition to the measurements transferred over the network, the dataset also contains measurements recorded by the data logger. The data logger recorded every reading by every query by the network until the logger ran out of memory. The resulting dataset contains two data files, data-log and data-net, which contained the measurements taken by the data logger and the network, respectively. Like a typical data frame, each column of the file contains a variable and each row, an observation. The dataset also includes meta-data which contained information about the location of the sensors (mote-location) and the mapping between the sample epochs and date-time (dates).

2.2 Data Cleaning

The data contained inconsistencies and anomalous measurements which had to be removed or corrected before further analysis. There were two data files: data-log and data-net. The data-log and data-net files contained readings recorded by the data logger and the network, respectively. To obtain a clean, combined dataset, I cleaned the two data files separately and combined them to form a single data file.

First, I removed duplicated rows and any observations that had missing values for all the measured variables. Next, I noticed that the date-time given in data-log was the same for all observations (2004-11-10 14:25:00),

which is nonsensical and also outside the range of dates of the experiment. I changed the time of the measurement using the dates file, which maps the sample epochs to date-time. For data-net, the sample epochs did not map to the same date-times as those given in the dates file. For a given epoch, the difference between the date-time given in data-net and dates was about 7 hours. To determine which epoch/date-time mapping to use, I compared the temperature time series for a single node for a single day. When I used the date-time given in dates, the temperature rose around sunrise, whereas using the date-time given in the net file showed that the temperature starts to rise sometime in the early afternoon. Since we expect the temperature to rise around sunrise, I decided to use the mapping given in dates.

Another inconsistency was the voltage value in data-net. The voltage values in data-net were in the 200s whereas those in data-log were between 0 to 3V. A comparison of the voltage values between the corresponding observations (i.e. same epoch and node combination) in the two files showed that the two voltage values had a linear relationship with almost perfect correlation. Using this linear relationship, I transformed the voltage values in voltage-net, which made the transformed values to be between 0 to 3V. Voltage was also associated with anomalous temperature readings. In data-log, voltage values of less than 2V corresponded to negative temperatures, which is unrealistic in the spring and summer months in Sonoma. In data-net, unrealistically high temperature values were observed about 30 percent of the time when the voltage was below 2.4V. Since the voltage level seemed highly correlated with correct readings of the temperature, I removed any observations whose voltage level fell below 2.4V as a conservative cutoff.

Furthermore, a significant portion of the observations had PAR values that were outside the range given in Table 1 of the paper. More specifically, 36 percent and 41 percent of the observations had incident PAR values greater than 2154 in data-log and data-net, respectively; 43 percent and 49 percent of observations had reflected PAR values less than 180 in data-log and data-net, respectively. I suspect the unit of measurement may be different between the data files and the paper. The observations that corresponded to exceedingly high PAR values did not seem to correlate with any of the other variables such as voltage, humidity, temperature, location of the sensor, sensor ID, or time of the day. Since I did not have any other information and was hesitant to remove a significant portion of the data, I kept these observations as they were. Furthermore, about 1.6 percent of the observations had incident PAR less than reflected PAR. Since such observations do not make any physical sense, I removed the observations for which incident PAR was less than reflected PAR.

After filtering for voltage and PAR as described above, data-log contained 68 epoch/node combinations that were duplicated but whose sensor measurements were different between them. 12,729 such combinations were found in data-net. For temperature and humidity, the variance of the measurements within the duplicated observations were small. For PAR values, however, the following node and epoch combinations showed high variance in data-log: (40, 36), (105, 7074), and (129, 7074), where the first and second elements are node ID and epoch, respectively. Similarly, (74,9441) showed a high variance in data-net. These four node/epoch combinations were manually removed. For the remaining duplicated observations, a random observation was chosen to represent a given node/epoch combination.

Finally, after checking that corresponding rows in data-log and data-net had the same sensor measurements, I merged the two files by node/epoch combination to form a combined dataset.

2.3 Data Exploration

To explore the data, I first looked at the distribution of values for each measured variable. As mentioned in Data Cleaning, I found that the values for incident and reflected PAR were much greater than what was described in the paper, which led me to suspect that the unit of measurement may be inconsistent between the paper and the data files. Similarly, looking at the distribution of other variables also revealed anomalous readings, which were then removed in the data cleaning stage.

Next, I explored how each variable behaved over the course of a day. For example, the humidity time series for May 7 showed that there was a dip in the humidity level from late morning to late afternoon, which could potentially reflect the changes in temperatures during that time (Fig 1). It also showed that there were measurements before 10 a.m. that fall much below the rest of the readings and led me to question whether some microclimatic phenomenon could explain this difference. Although I could not come up with an explanation for this specific example, looking at the time series data allowed me to discover interesting

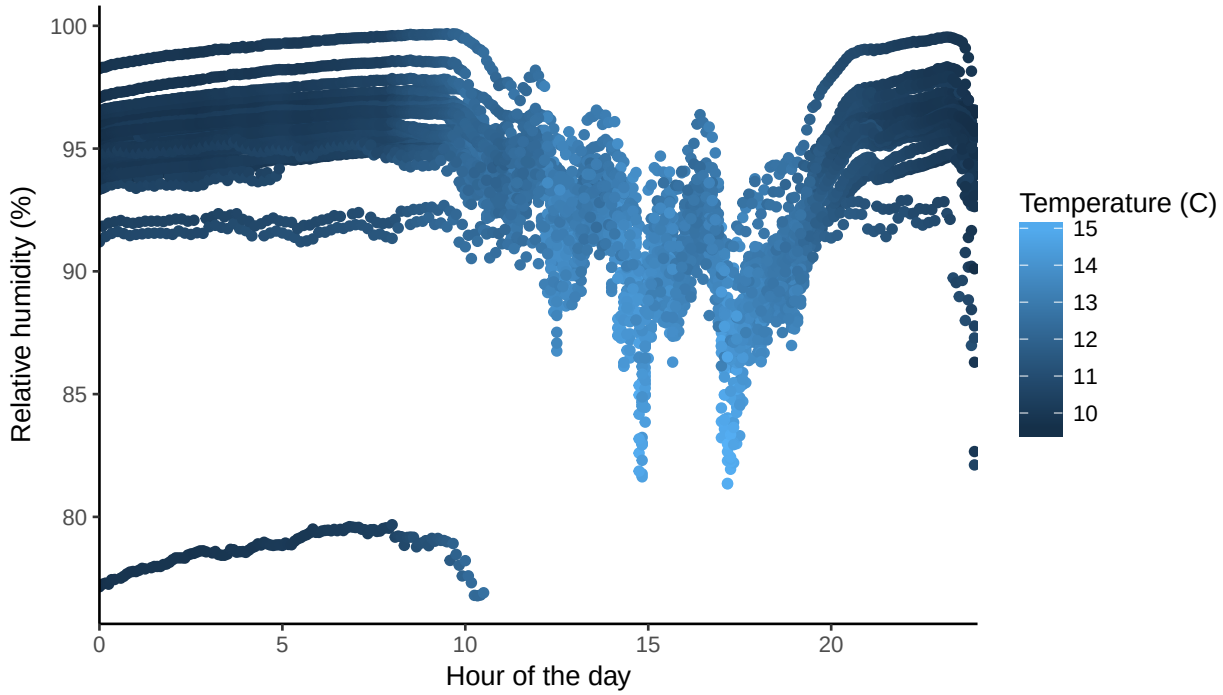

Figure 1: Relative humidity over time on May 7.

or unexpected trends in the data. Similarly, I looked at how each variable behaved over the length of the tree to find both expected and unexpected patterns, which will be discussed in Findings.

Next, I explored pairwise relationships between variables. For example, I made a scatter plot of temperature and humidity and found that they were inversely correlated (Fig 2). Although this trend was expected, it raised more questions about the relationship between humidity and temperature: Does this mean temperature and humidity should move in opposite directions throughout the day? Are there any periods during the day in which this is not true? A similar approach was taken to investigate the relationships between other variables, which led me to discover interesting trends discussed in Findings.

3 Graphical Critique

Figure 3a tries to show that the data points fall within an expected range and thus can be trusted for further analysis. Although the plots do show that the values fall within the normal range, they also raise further questions. First, it is unclear whether the distribution of PAR readings is actually bimodal. Because the majority of the data belong to the leftmost bin, it is difficult to see where the second mode is. On a related note, it is also curious to see why so many data points fall in the leftmost bin. Furthermore, a density plot may be more aesthetically pleasing since all the measured parameters are continuous.

Figure 3b shows the variation of each variable in each day and their general trend over the 44-day study period. I find these plots too busy. Selecting fewer days or sampling the days more sparsely, say every 5 days, could have made the plots easier to read while retaining enough information to show the overall trend. The incident PAR plot is particularly difficult to read as the outliers look like they form a solid line.

Figure 3c and 3d show how the variables behave along the length of the tree by combining all the measurements for a given height. These plots show that the sensors at every height reached practically all the possible values of humidity and temperature, suggesting that the amount of variation over time overwhelms the amount of variation over space. Again, I think the plots could have been less dense along the y-axis. One idea is clustering the height into 5-meter spacings (e.g. 5m, 10m, ..., 70m) and taking the average within

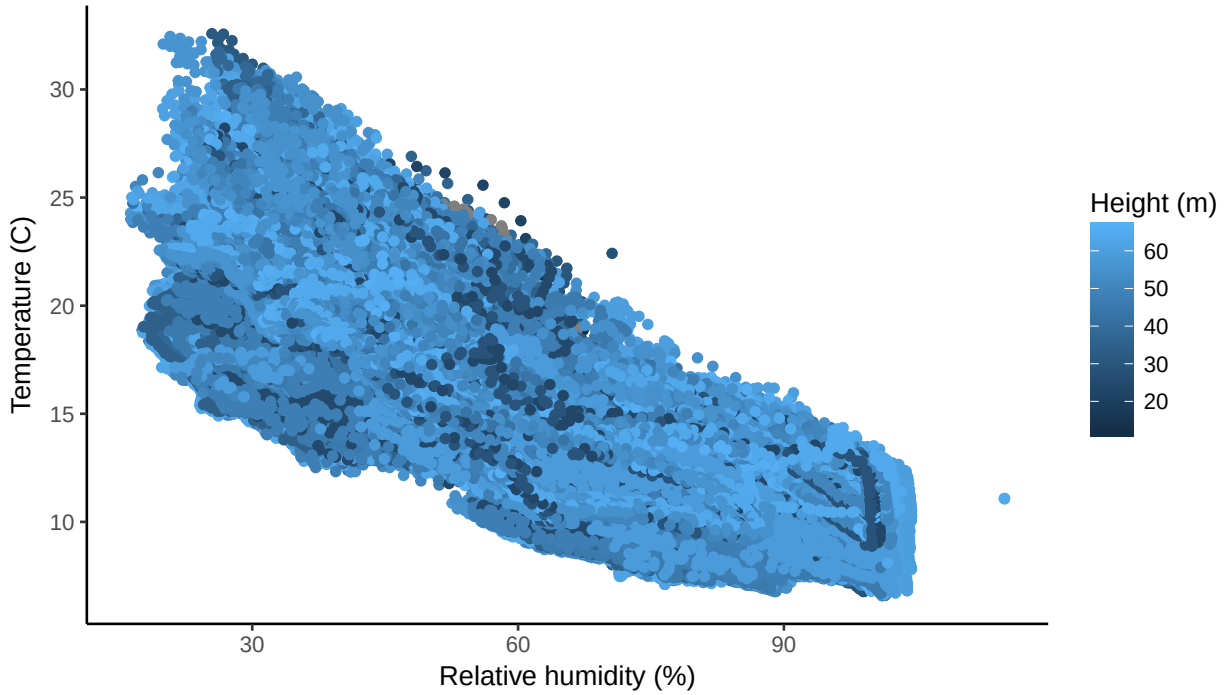

Figure 2: Relative humidity vs temperature

each cluster. Furthermore, Figures 3c and 3d provide redundant information and only one of them would have sufficed.

Figure 4 shows the day in the life of a redwood tree. The time series plots on the left part of the panel are informative in that they show not only the general trend throughout the day, but also the variation among the nodes at a particular time. For the time series of PAR values, however, it is unclear how the blue trend line was computed and whether the jaggedness of the trend line is informative at all. I would choose to add a smoother curve to show the general trend as the small jaggedness does not seem to provide much information. The right side of the panel shows the gradient of the parameters along the length of the tree at a single time point. I like these plots but I would add a legend indicating what the different color triangles mean. Also, it is unclear why the authors chose to fit a linear trend line for PAR gradients, whereas a quadratic trend line was used for the temperature and humidity gradients, especially since the linear curve seems like a poor fit for the PAR gradients.

4 Findings

4.1 First finding

Here, I explored the effect of the angular location of the sensors to the temperature measurements. Figure 3 shows the temperature changes throughout the day for sensors facing northeast and those facing west. The temperature readings were averaged over a 7-day period from May 1 to May 7 for each time point in the day. To minimize the effects of radial distance of the sensors to the temperature measurements, I only used sensors that were deployed at radial distance of 0.1m. Since this filtering step removed the sensors facing east, I used the sensors facing northeast as a proxy. From the figure, we can see both groups of sensors have a similar trend in the temperature changes throughout the day. The temperature rises from sunrise until late afternoon, after which the temperature gradually decreases. There are subtle differences, however. The rate at which the temperature changes during the day are similar for both groups but the sensors facing northeast measure higher temperatures on average than those facing west. This is sensible since the sun

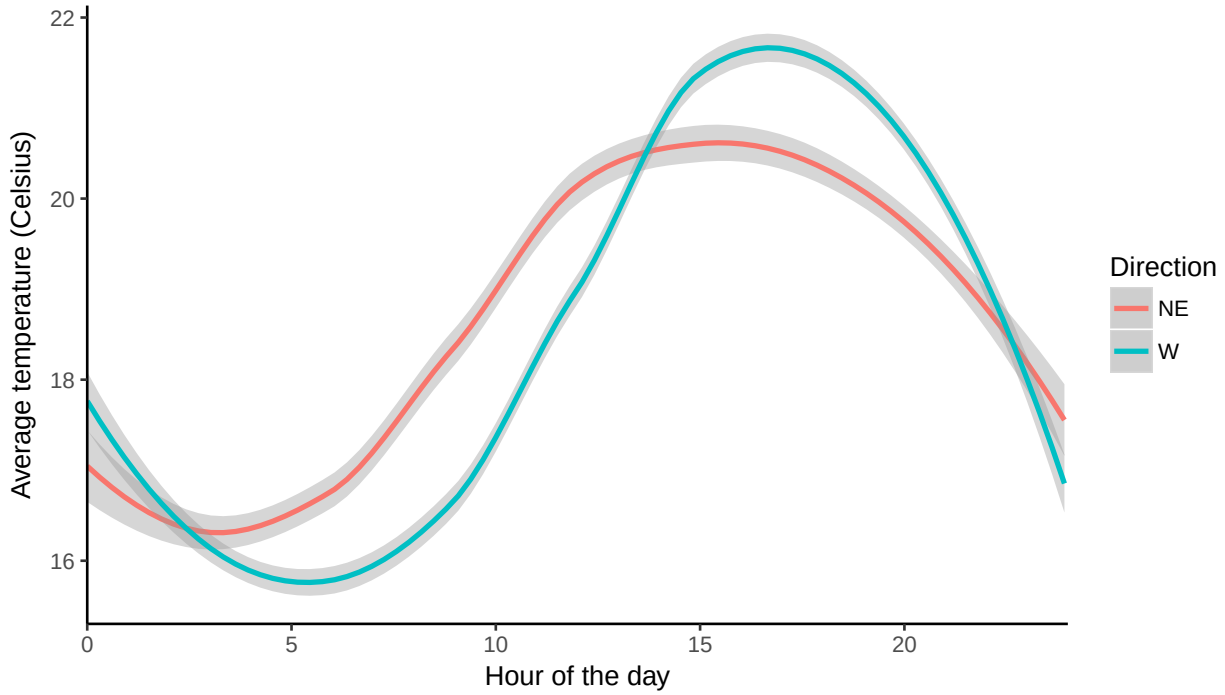

Figure 3: Temperature changes throughout the day, averaged over 7 days (May 1 - 7)

rises from the east so the sensors in the northeast get more direct sunlight than those facing west. Close to sunset time in the west, on the other hand, the average temperature for west-facing sensors are greater. This is particularly interesting since the median height of the west-facing sensors are about 10 meters below the northeast-facing sensors. So even at lower heights, the west side of the tree experiences higher temperatures.

4.2 Second finding

Figure 4 shows the relationship between humidity and temperature over the course of a day. Since temperature and humidity seemed to have an inverse relationship in Data Exploration, I investigated whether they also change in opposite directions over time. I averaged the temperature and humidity values of all nodes for a given time on May 2. Again, I only used sensors deployed at radial distance of 0.1m to see the environmental dynamics nearest to the tree. The measurements were then normalized to plot both curves in the same range. Surprisingly, Figure 4 shows that from sunrise to the early afternoon, humidity and temperature move in the same direction. This is opposite of what I expected; we expect higher temperatures to produce lower relative humidity since warmer air can hold more water. So the increase in relative humidity must be due to something other than temperature. One possibility is the transpiration process, which might overwhelm the effects of temperature changes. Figure 4 also generally agrees with the observation made by the authors of the paper who noted that movements in temperature and humidity do not necessarily correspond with each other during some periods in the day. After around 3pm, however, temperature and humidity do move in opposite directions as expected.

4.3 Third finding

Figure 5 shows the incident PAR levels over the course of a day and along the height of the tree on May 2. As mentioned in Data Cleaning, I was not sure of the unit of measurement for PAR so I normalized the data to show unit-less PAR values. Again, I only used the sensors at radial distance of 0.1m to investigate the microclimate nearest to the tree. We can see from the figure that in the early morning, light gets through to

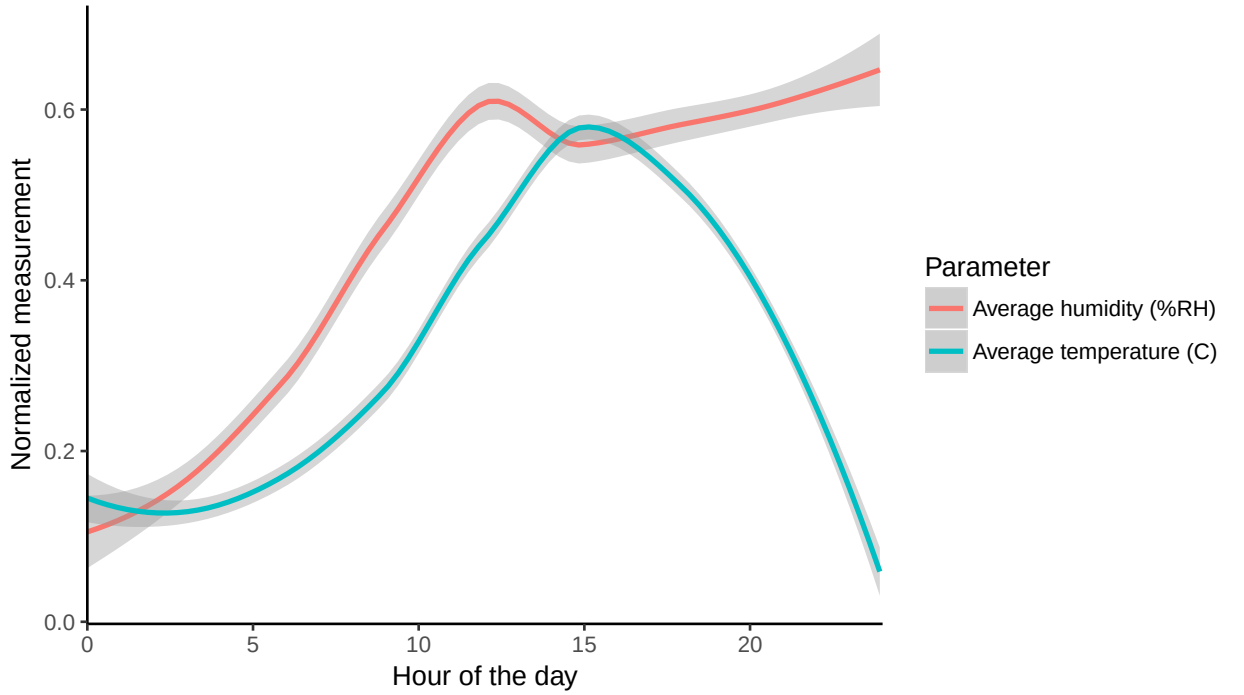

Figure 4: Temperature and humidity changes throughout the day on May 2

about 40m above ground. As the day progresses, the light travels farther down to the lower levels. Around noon, even the lowest sensors measure very high PAR values. It is interesting to see that the incident PAR level does not strictly correlate with the height of the sensors. For example, there are instances in which the higher sensors have lower incident PAR than lower nodes. This may be due to other factors such as wind moving the foliage. However, it is important to note the authors' mention of the sensitivity of the sensors to the orientation and their resulting fluctuations in PAR readings. Any conclusions drawn from PAR measurements, therefore, should be taken with a grain of salt.

5 Discussion

In this report, we saw that data obtained by wireless sensor networks can enable us to gain a greater understanding of environmental dynamics of a redwood tree. The data provided a dense set of measurements of climate parameters such as light, humidity, and temperature over both time and space. Although having a dataset of this size is very useful, we must tread carefully in analyzing the data. First, the data must be cleaned carefully. In this particular dataset, there were inconsistencies in voltage measurements, time records, duplicated observations, and measurements that were out of range of what was expected. While cleaning the data is important, we must also be careful not to "over-clean" and remove potentially interesting outliers that may reflect meaningful phenomena.

Although the large datasets are useful in finding trends in the data, it was also quite overwhelming to handle various dimensions of the data. I think we are very good at seeing patterns even when there are none, which may lead us to find patterns just by chance especially when there are so many data points. To make the dataset more manageable, I focused on analyzing a subset of days or a subset of sensors at a time. Furthermore, even though this dataset was small enough to load into R and perform various functions on it, larger datasets may require more work. For example, we may not be able to load the data all at once and may have to work on smaller subsets at a time; or we may not be able to use computational methods in a naive way and may have to resort to more sophisticated methods that can handle large datasets.

Another important lesson I learned was the necessity of domain knowledge in data analysis. For example,

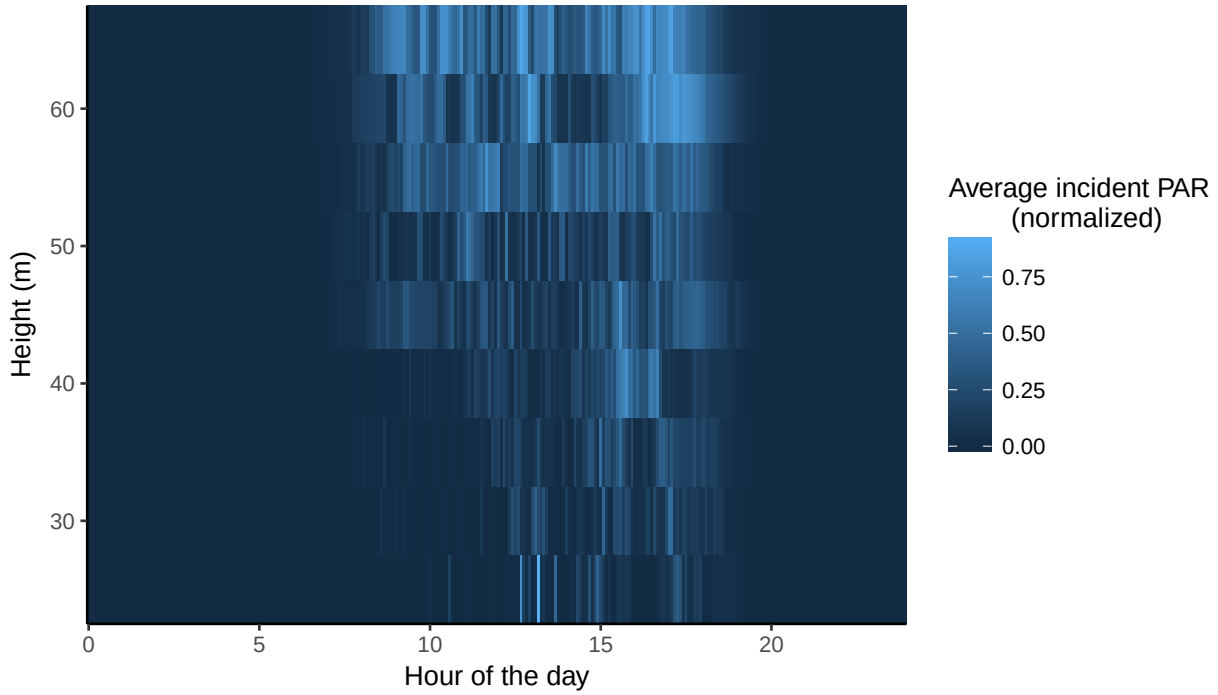

Figure 5: Incident PAR gradient over time and space

the range of PAR values in the data files did not match what was presented in the paper. Knowledge of PAR could have helped explain why the measurements looked strange and come up with possible ways to fix the problem. Also, my interpretation of the data was very much limited by my knowledge about the natural world. For example, I saw that the lower levels of the tree received just as much light as the higher parts. But I cannot say whether this is a meaningful observation that describes the real world or I am reading too much into noisy data. Overall, the study highlighted the importance of domain knowledge and of careful data cleaning.

6 Conclusion

In this report, we saw that wireless sensor networks can enable us to monitor the world at a macroscopic level by providing temporal and spatial measurements of various climate parameters for a redwood tree. In particular, we saw that temperature measurements are sensitive to the angular location of the sensors; humidity and temperature do not necessarily move in opposite directions over the course of the day; and that even the lowest parts of the tree receive a full range of incident PAR during the day. This study highlights how such data have a potential to provide a greater understanding of the natural world at a macroscopic level.

References

[1] Tolle, Gilman, et al. "A macroscope in the redwoods." Proceedings of the 3rd international conference on Embedded networked sensor systems. ACM, 2005.