Lab 1 - Redwood Data, Stat 215A, Fall 2021

Chengzhong Ye

September 16, 2021

1 Introduction

Tolle et al.[1] constructed a monitoring device called "macroscope" on Californian redwood trees, which is essentially a system of sensor node networks deployed on various locations on the tree that record a group of climate variables over an extended period of time. This provided researchers an unprecedented opportunity to study the microclimate ("a local set of atmospheric conditions that differ from those in the surrounding areas" [2]) around a redwood tree. Previous studies of the the microclimate were only able to obtain "snapshots" of the climate. While with this network of sensor nodes that continuously measures climate variables, we are able to obtain a full spatio-temporal landscape of the microclimate on the redwood tree. This will help us better understand the environmental dynamics around the trees and how they interact with them.

However, since this was the first attempt of setting up such a system, the recording process turned out to be problematic. Sensor were found to report abnormal values. And due to battery life and storage limit, many nodes stopped recording at different time points before the planned end date. These problems has led to poor data quality and present challenges for data analysis. In this report, we will first briefly describe the dataset. Then we will discuss exploratory procedures to identify issues in the data and clean the data. We showcase how data cleaning can affect the result of analysis. And we make use of the spatio-temporal information of the recordings to demonstrate that such data is indeed helpful for studying the microclimate surrounding redwood trees.

2 Data

This is a time-series dataset that includes measurements of four climate variables: temperature, humidity, incident PAR and reflected PAR (photosynthetically active radiation, which represents energy available for photosynthesis and also indicates the overall strength of sunlight). The data were recorded with a time interval of 5 minutes from April 27th to June 10th 2004, 44 days in total. And they were measured by 73 sensor nodes located on different heights of two redwood trees. The spatial distance between nodes are about 2 meters. The nodes are also placed on different sides of the tree, with the majority on the west side and some on the east side. This resulted in totally 416,036 measurements of the four variables respectively.

2.1 Data Collection

The data were collected by specially designed sensor nodes. Each sensor node is hat-shaped. One the top there is the sensor for measuring incident PAR and therefore receiving sunlight from above. At the bottom we have the sensors for temperature, humidity and reflected PAR. This is to prevent the sensor from being affected by direct sunlight and other environmental disturbances such as wind and rain. The temperature and humidity were measured by the Sensirion SHT11 digital sensor in Celsius ($\pm$ 0.5°C) and relative humidity ($\pm$ 3.5%). The incident and reflected PAR were measured by two Hamamatsu S1087 photodiodes in μmol/m/s. These sensor nodes form a network and the data were transmitted to a recording system. Also each sensor has as data logger that stores the measurement data locally. Finally the data were retrieved in both ways.

2.2 Data Cleaning

The cleaning procedures were first carried out on the data from the network (we call it net data below) and from the logger (we call it log data below) separately. After that the data were combined for the analyses. In this cleaning process, we in principle aim to identify malfunctioning nodes by abnormalties in data, and then look into their recordings and decide whether to remove all or part of its data.

First, there are plenty of NA values in the data. Instead of removing them directly, we want to examine if these missing values are restricted to certain sensors, which could indicate whether their recordings are unreliable. In the log data, we found all the NA's come from three nodes: 15, 122 and 128.

nodeid	proportion of NA
15	0.6804608
122	0.6547046
128	0.6197822

In the net data, we found all the NAs are from node 1.

nodeid	proportion of NA
122	1

Therefore, we look at temporal change of recordings by these three nodes.

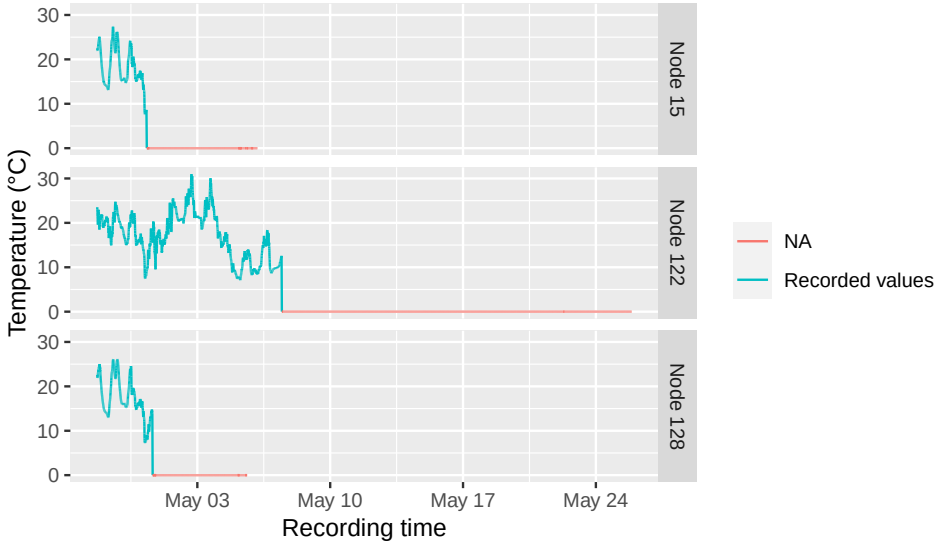

Figure 1: Temperature recordings of nodes 15, 122 and 128

Fig.1 shows that the missing values were all generated after a certain time point and the recorded data before that point look fine (here we only showed temperature measurements, but the same can be observed for others). So we only remove the NA's. In the net data, all the NA values come from node 1 and it has no recordings at all. Therefore we also just remove all the NA's.

Next we simply remove abnormal values that violates physical laws, namely entries with temperature measurement larger than 100 and humidity larger than 100 or smaller than 0.

According to the original paper, abnormal voltage recording may indicate unreliable measurements in the corresponding entry. Motivated by this, we examine how the voltage changes by time for each node in the log

data. As shown in Fig.2, we found some extremely low recordings. So we check the measurements in the corresponding nodes. It turned out that except for one node with abnormal incident PAR recordings, other measurements are within normal range. Thus, we only remove data from this node. We also find incident PAR recordings from another node with shifted baseline that is not related to voltage abnormality. Therefore, we remove data from both nodes.

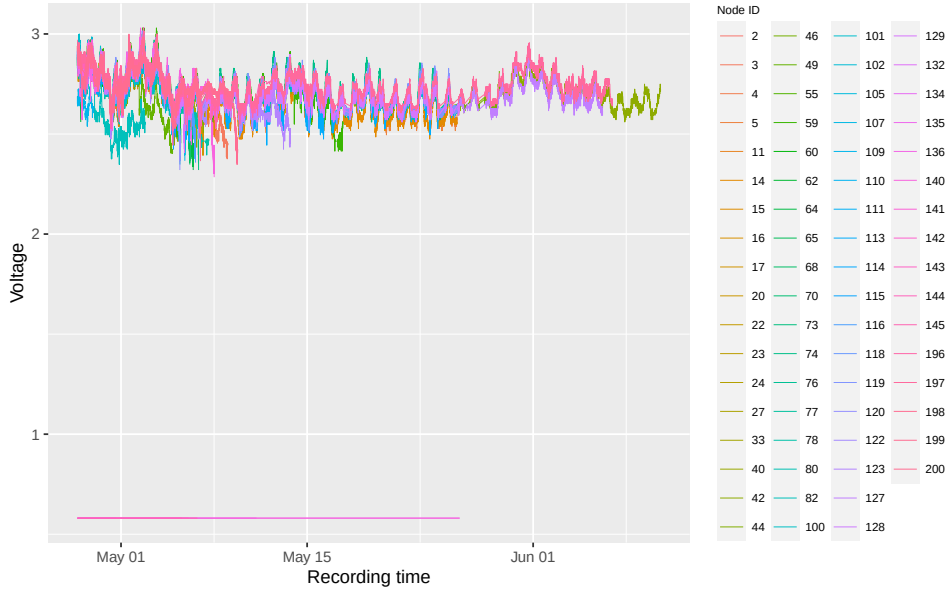

Figure 2: Voltage recordings of each node in the log data

We then do the same thing for the net data. As shown in Fig.3, we found a node with consistently very high voltage and several nodes with voltage increasing rapidly after a few epochs.

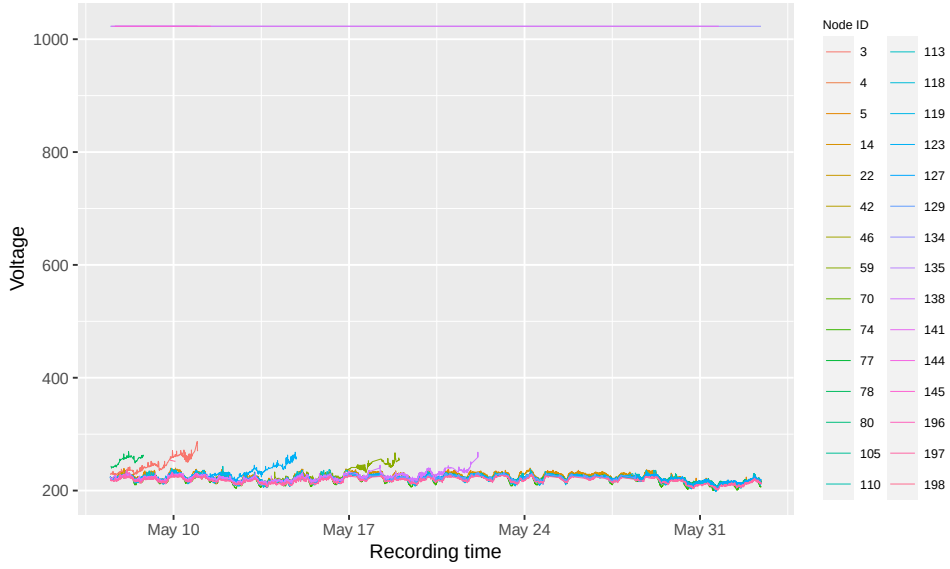

Figure 3: Voltage recordings of each node in the net data

When looking into measurements from these nodes, we observed the corresponding abnormalities in the temperature recordings (Fig.4). So we removed them.

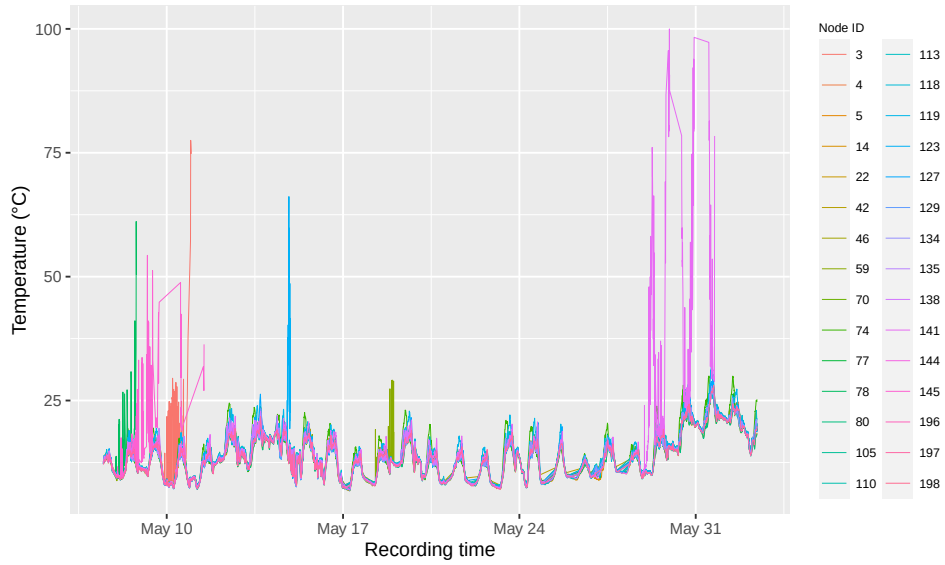

Figure 4: Incident PAR recordings of abnormal-voltage nodes in the net data

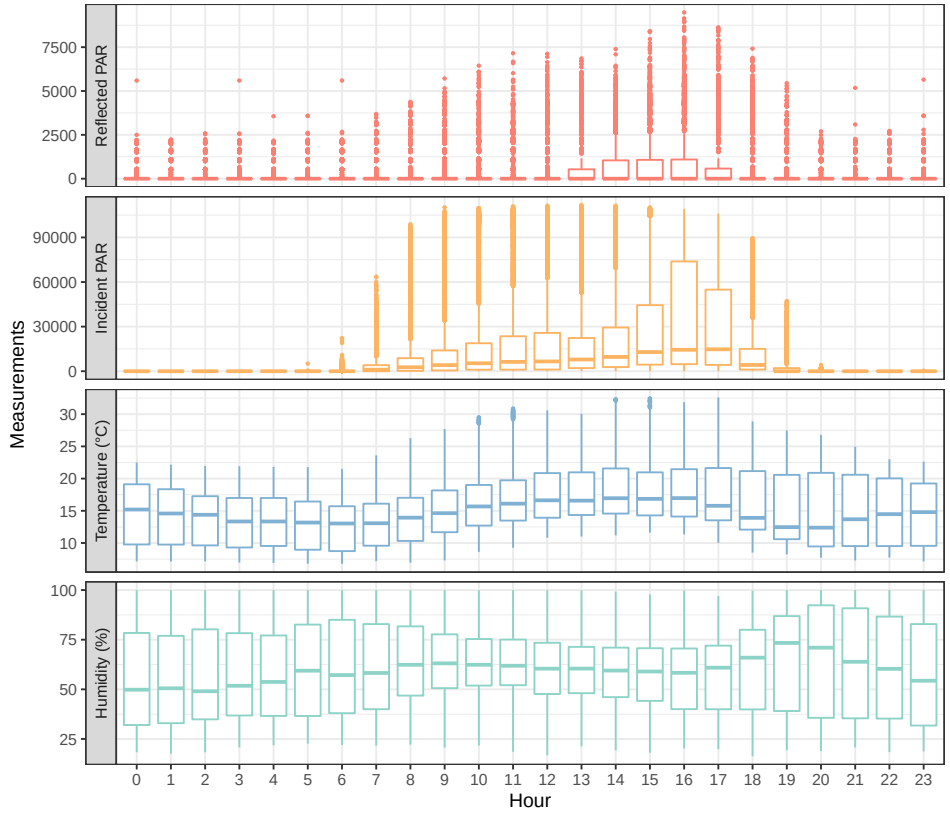

Figure 5: Four climate measurements summarized by 24 hours in each day

Then we examined duplication in the data. Each node should only have one recordings in one epoch. But we found about 4% of log data and 10% of net data are duplicated. However, the humidity and temperature measurements are very close between duplicated entries with only 38 pairs (~0.1%) of duplicated recordings having standard deviation larger than 1 in at least one type of measurements. We found most of these

discrepancies come from the reflected PAR measurements. And they come in the form that in one pair of duplicated entries, the reflected PAR values is zero in one entry and high in the other. Here we remove the entries with zero values. For other duplicated data, we ignore the difference and just retain the first recordings.

Then we removed nodes with very short active time, namely those stopped recording before May 1st. We also removed data on the first day April 27th since the recording only started from late afternoon.

Finally we combine the cleaned log and net data. If both the network and the logger have recordings from a certain node in a certain epoch, we use the log data.

2.3 Data Exploration

There are two trees in the data. So we split the data by trees and restricted the following analyses only to the interior tree. The same analyses were also done using data collected from the edge tree and very similar results were obtained. Due to limited length we do not show it here.

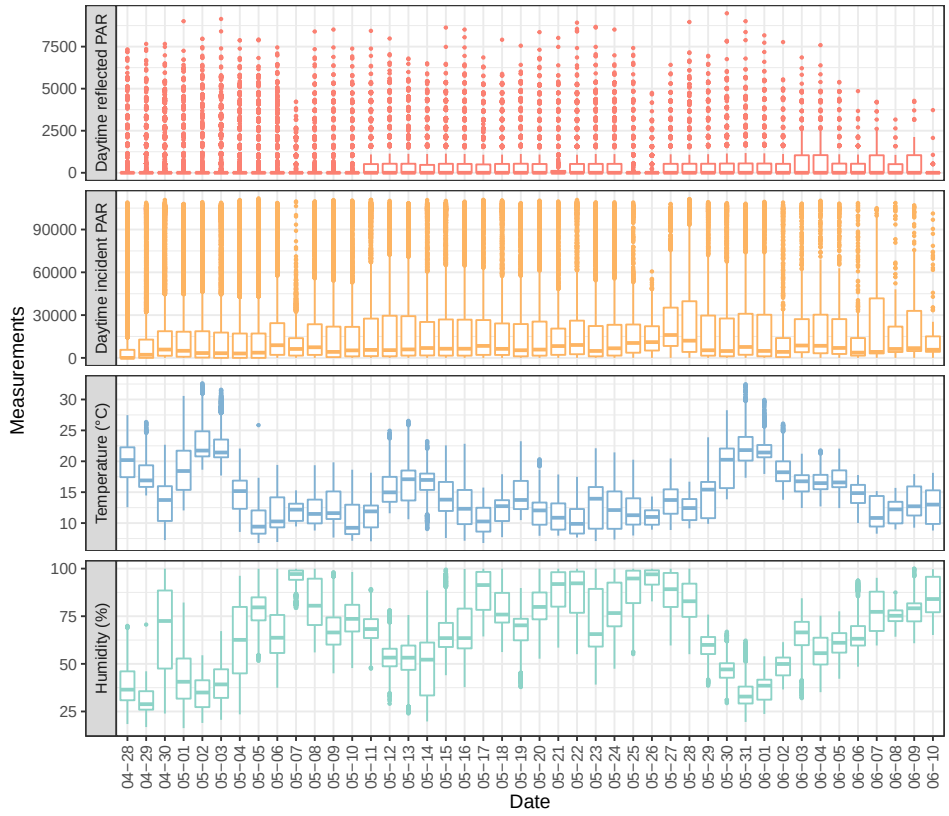

Figure 6: Four climate measurements summarized by date

In the paper, the authors summarized the temporal pattern of the measurements by the date. But we claim that the within-day 24-hour pattern is also worth exploring. Here we use boxplot grouped by the hours in a day to summarize data across days and sensors (Fig. 5). We found that the incident PAR recordings are generally zero during night time, then start to increase from 6 in the morning, then reach the peak around 16 in the afternoon and then rapidly decrease to 0. Given the majority of the sensors are west-facing, this pattern can be expected. The reflected PAR measurements showed similar pattern while the values are relatively lower and more noisy. Interestingly, the temperature measurements shares a similar pattern as the incident PAR. This suggests that direct sunlight on the tree may largely determine the temperature of the particular region on the tree. We further investigate this relationship a later section. The humidity

measurements do not show a clear trend throughout the 24 hours.

Then we summarized the data by different days (Fig. 6). PAR measurements have very large zero fraction, which is partly due to zero sunlight at night. Therefore, here we only included daytime measurements (6AM to 6PM) to focus on the more meaningful values and to help with visualization. It can be observed that the variation of PAR measurements across days are minor, while the day-to-day differences of temperature and humidity are prominent. In a later section, we investigate daily weather differences in further detail. Interestingly, we found that the humidity and temperature measurement seem to have a negative association. This could be explained as in the paper that colder air has lower capacity of holding moisture. Therefore the same absolute water content will be converted to higher relative humidity when the temperature is lower.

We also examined the spatial pattern of the measurements by grouping data by sensors (Fig. 7). We found that the PAR values are clearly associated with the height of the sensors. Highly positioned sensors tend to have much higher PAR measurements than those at the bottom. However, it seems that the spatial variation for temperature and humidity are not prominent. By further investigation, we found this is not the case. And the variation is actually masked due to poor data quality, which we will discuss later in the report.

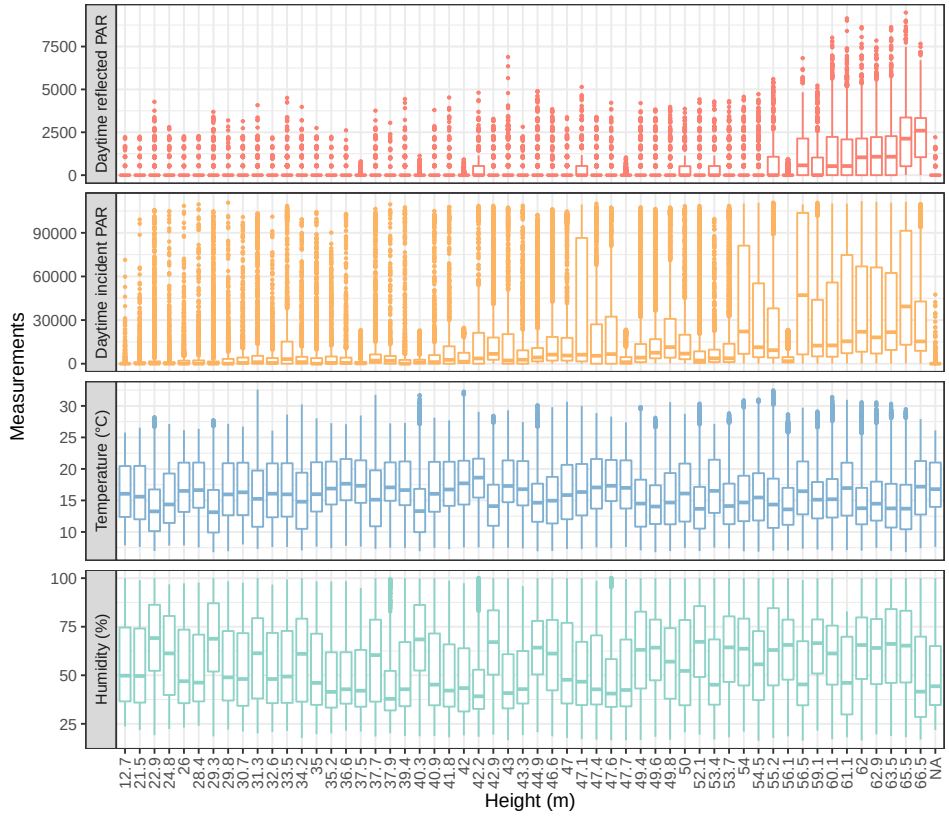

Figure 7: Four climate measurements summarized by height of the sensors

As mentioned in the paper, the number of successful recordings decreases as time goes on due to died batteries, etc. Here we want to see how many recordings are made on each days in our data. As shown in Fig. 8, the number of recordings began to drop at an approximately linear rate after May 6th for sensors. This extremely uneven temporal distribution of the amount of data is likely to introduce bias if we pool across different sensors in our analysis. Therefore we explored and used data within a certain range of time for some of the analysis below.

Figure 8: Number of successful recordings on each day

2.4 Reality Check

First, we checked if the cleaned data is in line with our common sense, the weather measurements are within reasonable range. For example, temperature over 50 degree Celsius is very rare and we do not expect to see such extreme values in the data. The temperature and humidity values change in a continuous manner across the time. As the air absorbing or emitting heat or moisture is a gradual physical process. The temperature is typically the highest in the afternoon and the lowest around dawn. Sunlight from the west reaches the peak in the afternoon and zero in the nights. Higher positions on a tree is likely to receive more sunlight and we do observe a trend of higher sensors having higher PAR measurements.

Then we performed external validation. We downloaded the historical temperature recordings at the Sonoma county in May 2004 from the National Center for Environmental Information [3]. And we compared the measurements from the redwood sensors to this data. As shown in Fig. 9, we found that the highest daily temperature recordings showed a strong correlation with highest daily sensor measurements with Spearman's rho = 0.77. This suggests that our data is reliable because the temperature pattern across different days agrees with external data. Although the dots are slightly shifted to the bottom right, meaning that the values in historical recording tend to be higher. But such bias can be explained by the distance between the recording station and the redwood tree in the experiment.

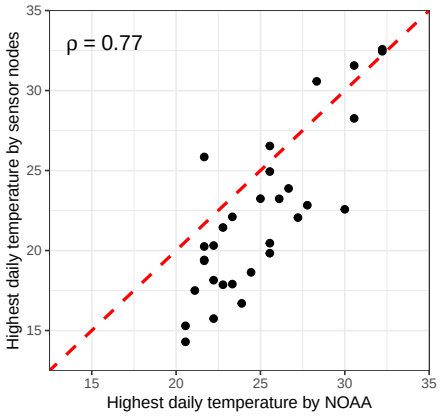

Figure 9: Highest daily temperature measurement in line with external data

3 Graphical Critique

Figure 3: Looking at the histogram at the first row. The PAR measurements showed two features: skewed distribution and large zero fraction. This has made the histogram and the following boxplots not very

informative. One possible modification is to perform Log transformation to the PAR values to deal with the skewness. Nevertheless, we have tried the approach and it is not very helpful. Another idea is to focus only on the daytime PAR values because PARs are expected to be zero during night time and these recordings are therefore less meaningful. And this is what we used to make the previous exploration graphs and it did improve the visualization.

Also, the two lower rows describing the value x height relationship is a bit repetitive and it may lead to confusion about the difference between the two rows. Also the mean subtraction approach may not be a good option for visualization because it introduces too much variation, making the boxplots too compact and uninformative.

Aesthetically, the text could be made a bit larger to be more visible. And we can use different color to denote different types measurements to convey more visual information.

Figure 4: First of all, this figure is not labeled clearly. The two figures above uses differently colored line to present data for each node and two figures below uses uniformly colored dots to represent data for all nodes and a line to show the overall trend. None of this elements are clearly labeled within the figure, making it difficult to interpret without referring to the main text. The text above actually denotes the vertical line showing the time point used to make the height figures on the right, but given the position it can be mistaken as the title and cause confusions.

The figure mainly wanted to show the change pattern of the four types of measurements within one day and the relation between each measurement and height. It does illustrate these points. But only showing data of one single day or one single time point can hardly be considered representative for the whole picture. It would be better to summarize across different days to show the pattern or relation observed are not restricted to this particular day.

Aesthetically, the text could be made a bit larger to be more visible. And the scale of axis can be improved for the right column of plots to make the relation clearer.

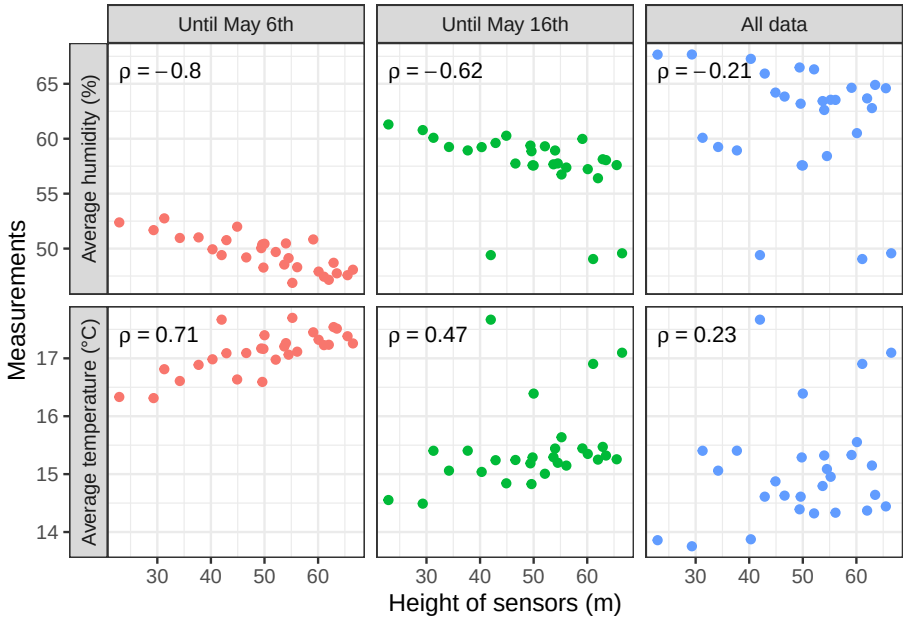

Figure 10: Correlation of average temperature and humidity vs height of the sensor in different time frame

4 Findings

4.1 First finding

Previously we found that unlike PARs, temperature and humidity do not show a clear correlation with the height of the sensors. A plausible explanation is that the temporal variation of temperature and humidity is more prominent than the spatial variation and the averaging step masked their correlation with height. However, after some investigation, we found that it is more likely due to the total recording time of the sensors being vastly different. As shown in the data exploration section, sensors stopped recording at different time. For example, sensor 3 only recorded data until May 9th but sensor 42 recorded data until Jun 10th. This means for the two sensors we are taking the average of temperature across two different periods of time. This additional fraction of temperature recordings (May 10th to Jun 10th) for sensor 42 will incur a large difference in the average value that masks difference associated with height.

To illustrate this, we chose three time points as cut-offs: May 6th, May 16th and June 10th (the final day) and calculated the average temperature and humidity recorded by each sensor on the interior tree. On days before May 6th, most of the sensors are working and therefore the temperature and humidity are averaged across the same period of time for most sensors. While part of the sensors stopped recording before May 16th and all sensors stopped before June 10th. Therefore the active time of the sensors are more heterogeneous within these two time frames, adding larger noise into the average values.

In the Fig.10, we can see that the correlation between average temperature measurement and height is very high using data until May 6th (Spearman' rho = 0.71). It drops to 0.47 using data before May 16th and becomes mostly unnoticeable with all the days (0.23). And a similar pattern is observed for the humidity measurements. This has shown that the average temperature and humidity do have strong correlation with heights.

4.2 Second finding

In this section, we try to classify and characterize different days by the four types of measurements. Now we summarize the temperature, humidity, incident PAR and reflected PAR in each day by averaging across different nodes. We included days until May 7th to verify the statement in the original paper that May 7th has a distinct foggy weather.

A hierarchical clustering is performed using `hclust()` function in `R` with default settings based on the four measurements. On the heat map on the left of Fig.11, we can see the days are categorized into three groups: hot and dry days, cold and humid days and one day with extremely high humidity and very low sunlight, which is May 7th. This shows are days can be categorized by the four types of measurements of weather and confirms the statement in the paper. To show how much these groups are distinct from each other, we performed a principal component analysis to visualize the distances in two dimension. Marked by different colors, on the right plot of Fig.11 we can see the four groups of days are visibly distinct on the space spanned by first two principal components. This overall has shown that such "macroscope" system can be used to record and characterize climate change in redwood forests.

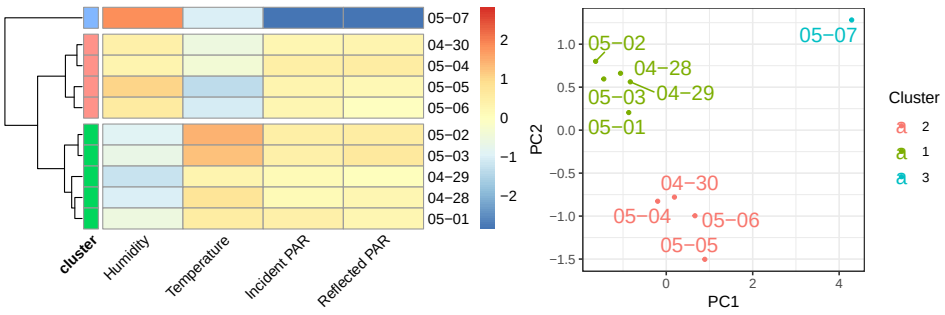

Figure 11: Days can be categorized based on measurements of the four climate variables

4.3 Third finding

In the exploration section we found that the pattern of temperature throughout 24 hours in a day coincides with the level of incident PAR. Here we take a further look into this phenomenon. Because the pattern of sunlight in a day is strongly associated with the orientation, we separate sensors into west-side and east-side ones based on their direction. As we all know, the sun is positioned at the east in the morning and at the west in the afternoon. Therefore east-side sensors are expected to have peak incident PAR recordings in the morning while west-side ones should have peak values in the afternoon. This can indeed be found in our data. In the first row of Fig. 12, we found that west-facing sensors typically record highest incident PAR values around 4PM and east-facing ones have typical peak values around 11AM, which is inline with our common sense. We then look into the temperature measurements in these two groups. We found that, in both groups, the hourly temperature is strongly associated with the level of incident PAR. And the calculated Spearman's rho turned out to be around 0.8, as shown in the bottom row.

The sensors placed on different sides of tree are still closely positioned (<100m). Therefore they are expected to share the same baseline "macro-climate". But the patterns of their temperature change are distinct and associated with sunlight. This suggests that the direct sunlight may have a strong influence on micro-climate on a redwood tree. This shows that such sensor system does provided useful information to help us study the climate condition around redwood trees at a finer resolution.

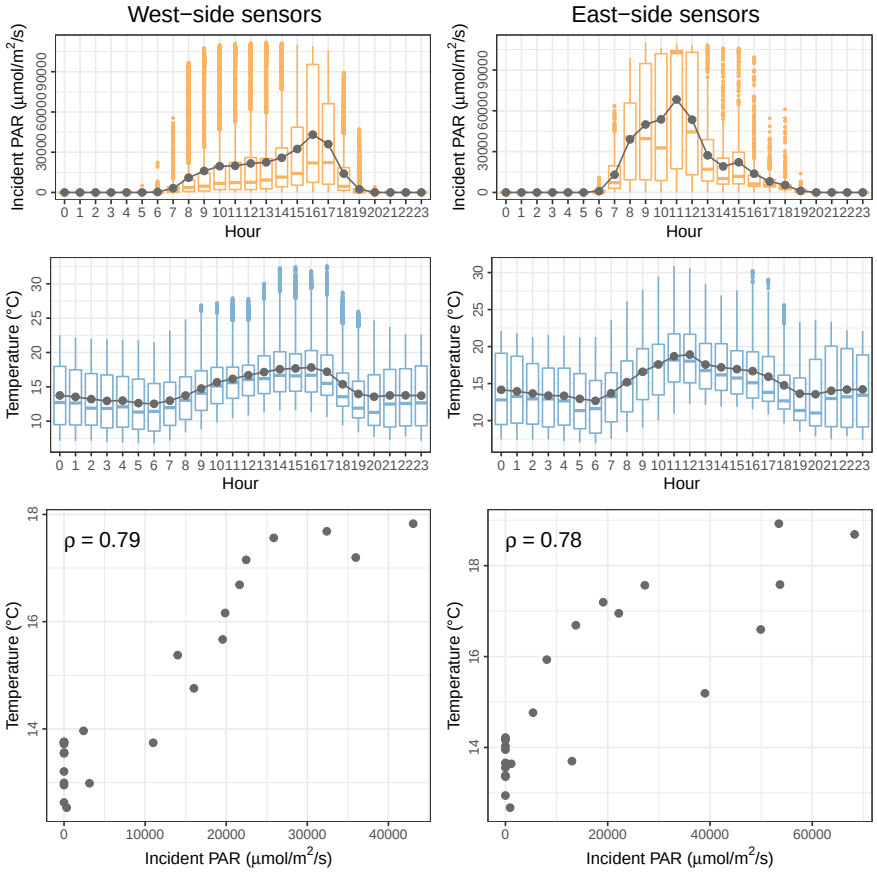

Figure 12: Hourly temperature and incident PAR are highly correlated on both the west and east sides of the redwood tree

4.4 Stability Check

In the first finding , we chose two cut-off dates, May 6th and May 16th, for the analysis. Here we try to add a bit more noise by postponing each of the cut-off date. We redo the analysis using May 10th and May

24th instead. In Fig. 13, we can see that the correlations overall did drop after we included more noisy data. But we still see that by restricting the time frame to when more sensors are active can make the association clearer.

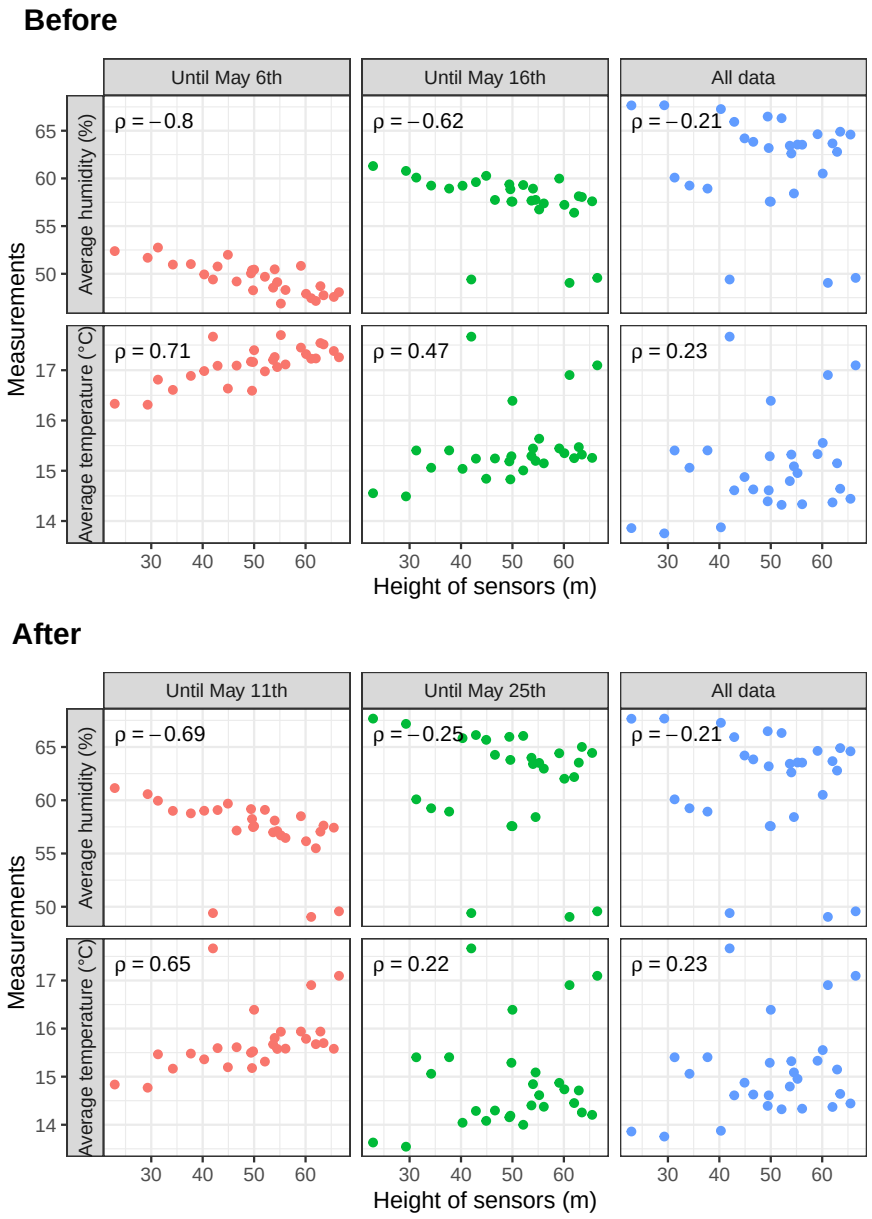

Figure 13: Perturbation of the first finding by changing cut-off dates

5 Discussion

One major problem in this dataset is the extremely uneven distribution of recordings across time and space. The active time frame varies significantly across different sensors. Therefore the amount of "complete" data is relatively small. For example, if we want to do some analysis across different days, only the data before May 7th have measurements from most sensors, which is about 1/4 of the whole experiment time span. If the recording is more stable, we could investigate the change of climate in a longer period of time. And it may also help with investigation of spatial variation because the nodes will have more comparable recordings.

Now we summarize this data analysis procedure by the veridical data science framework. The thing we want to study here is the micro-climate around a redwood tree. This is in the first realm: the reality. We collected time-series data of four climate measurements: temperature, humidity, incident PAR and reflected PAR. These data can only describe a slice of the reality as the reality is infinitely complex. And we expect these samples to be representative of the population and can be reflect some aspects of the micro-climate that are of interest. This collected data is in the second realm: representation of reality. And all the methods and analysis procedures fall into the third realm: mental construct. For example, the correlations and categories found in our analysis only exist in our mind and we use it as a medium to perceive the data. And one way to validate these mental constructs is to apply it to future data and see how well it can predict unseen data, which is the predictability principle. And if the models and algorithm can well characterize future data, which is related to future reality, we can then make decisions based on the models.

Data visualization is a main topic of this lab. I would like to consider it as a further reduced representation of the data, which is a representation of the reality. Because we are usually not making assumptions for visualization, it does not quite fall into the third realm. A good visualization should be intuitive, succint and highlight the aspects of interest in data, and therefore, reality.

6 Conclusion

Overall, we found that this system of sensor networks does provide meaningful data that can be used to investigate the micro-climate on redwood trees. With the spatial-temporal recording of four climate variables, we were able to obtain some interesting findings. The spatial information has been shown useful as all of the four weather measurements have a clear association with the height of the sensor. We also found that different side of the redwood tree tends to have distinct pattern of micro-climate. Also, this data enabled us to investigate the temporal change of weather around the redwood tree at a high resolution.

7 Academic honesty statement

Here I make the truthful statements: I myself designed and performed all the data analysis procedures presented in this report. I myself wrote all the texts and produced all the figures in this report. I have included and documented all the procedures in the workflow and the results can be fully reproduced. Wherever I included the work from others, I cited the sources.

I think academic honesty is an essential prerequisite for conducting any form of research. First, it is important that we take responsibility for our research results. Because science is built upon collaboration. Your results will be the foundation of other people's research. A dishonest or unreproducible research will possibly lead to a cascade of false results. Also, it is important that your work is original. Plagiarizing is not contributing anything new to the scientific community and it is disrespectful to the original author as you are taking their credit. Therefore, we should always keep our research honest, transparent and reproducible.

8 Bibliography

1. Tolle, G., Polastre, J., Szewczyk, R., Culler, D., Turner, N., Tu, K., ... & Hong, W. (2005). A macroscope in the redwoods. In Proceedings of the 3rd international conference on Embedded networked sensor systems (51-63).
2. Wikipedia: Microclimate. https://en.wikipedia.org/wiki/Microclimate
3. National Center for Environmental Information. https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00023213/detail