# Lab 1 - Redwood Data, Stat 215A, Fall 2018

Linqing(Waverly) Wei

September 13, 2018

## 1 Introduction

The emergence of new wireless sensor networks sheds light on the microclimatic analysis in the natural world. The sensor networks automate the data collection process, and more importantly, it adds depth and dimensionality to human's understanding of the surrounding world. The study is conducted on a coastal redwood canopy, a seemingly static and constant subject. However, with the assistance of "macroscope," the microclimatic system surrounding the redwood tree reveals the underlying dynamics and complexity. This analysis aims to explore the internal relationship of the microclimate system using the massive data obtained from "Macroscope." Since the microclimate system surrounding the redwood tree is complex in nature, each part of exploration would take both temporal and spatial components into consideration thoroughly before reaching to a consluion.

## 2 The Data

The redwood datasets contain 11 variables: result_time, epoch(time series), nodeid,parent(network strucutre), voltage,depth,humidity, humid_temp(temperature),humid_adj( no specific definition),hamatop(incident PAR), hamabot(reflected PAR). PAR denotes photosynthetically active solar radiation. redwood_all_orig contains 416036 observations, redwood_log_orig contains 301056 observations,and redwood_net_orig contains 114980 observations.

The "dates" dataset contains only temporal data with 5 variables:number, day, date,time, datetime, and 13000 observations.

The mote_location dataset contains spatial information with 5 variables: ID (nodeid),Height, Direc (E ESE NE NW S SW W WNW WSW), Dist, Tree (interior/edge)

### 2.1 Data Collection

The installation is set up on a 70-meter tall coastal redwood tree during a 44-day collection period. Data are captured every 5 minutes and every 2 meters in space. The sensors are placed 15m from the bottom to 70m, west side of the tree, which provides a buffering effect. The sensors are also very close to the tree, with a radial distance of 0.1 to 1m.

### 2.2 Data Cleaning

The outlier rejection session in the paper is not quite accuarte. When plotting voltage against temperature, within normal temperature range, voltage falls within the range of 2 to 3. When narrowing down to that voltage range, all normal temperature data points lie in voltage 2.2 to 3.0. Therfore, 2.4 to 3 shouldn't be used as the rejection criteria. I used 2.2 to 3.0 instead.

Fig 1(a).Outlier Rejection based on Voltage vs.Temperature:Overview

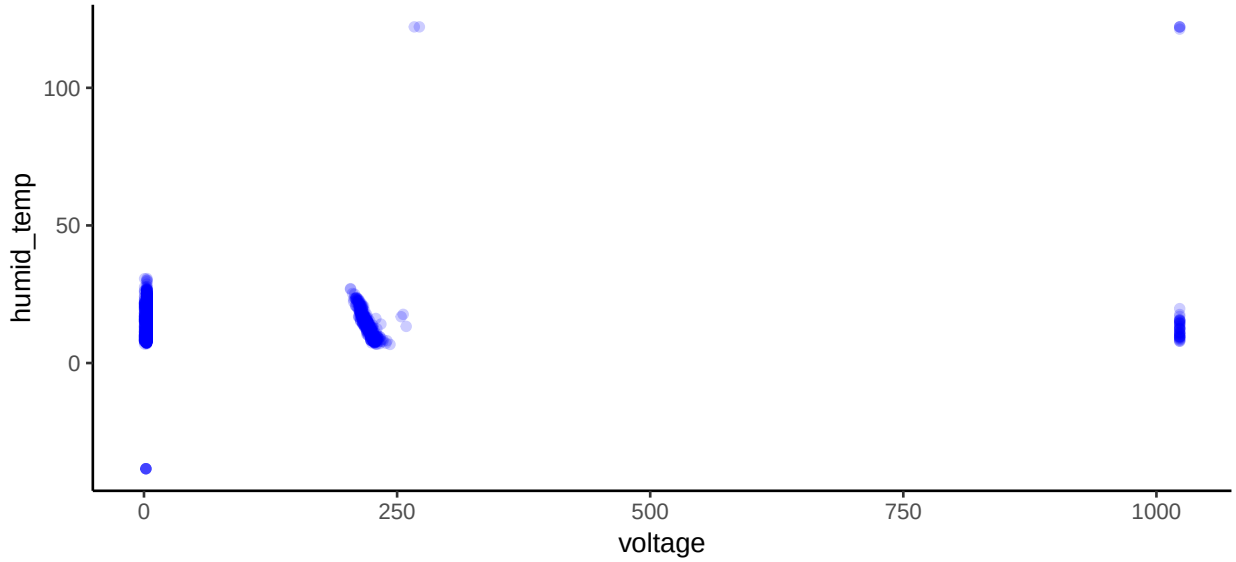
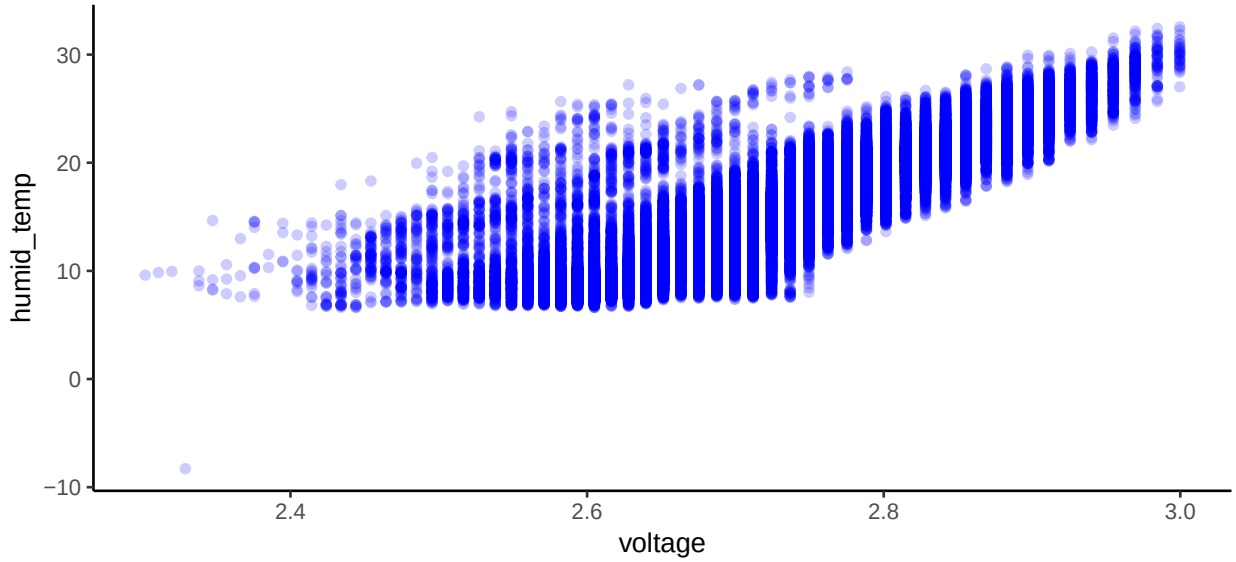



Fig 1(b). Outlier Rejection based on Voltage vs.Temperature:Zoom in




Then I removed outliers from humidity column since humidity cannot exceed 100%.Another point to notice is the incident PAR and reflected PAR. In general, reflected PAR is less than incident PAR, however, whether or not using that inequality as a filtering criteria worths more discussion. If reflected PAR is larger than incident PAR, then very likely the bottom sensor receives radiation from outside of the microclimate system. After going back to the paper, I noticed that the paper specifically states that the sensors are placed extremely close to the tree to control for the out-of-system noises. Therefore, I removed all the observations for which reflected PAR is larger than incident PAR.

## 2.3 Data Exploration

### 2.3.1 Explore Three Redwood Datasets:net,log,all

The total number of rows after correctly combining net and log datasets should be 341854 (merging is conducted based on the tuple epoch and nodeid), which means the current red_wood_all dataset has redundancy.

Table 1: Example Duplicates

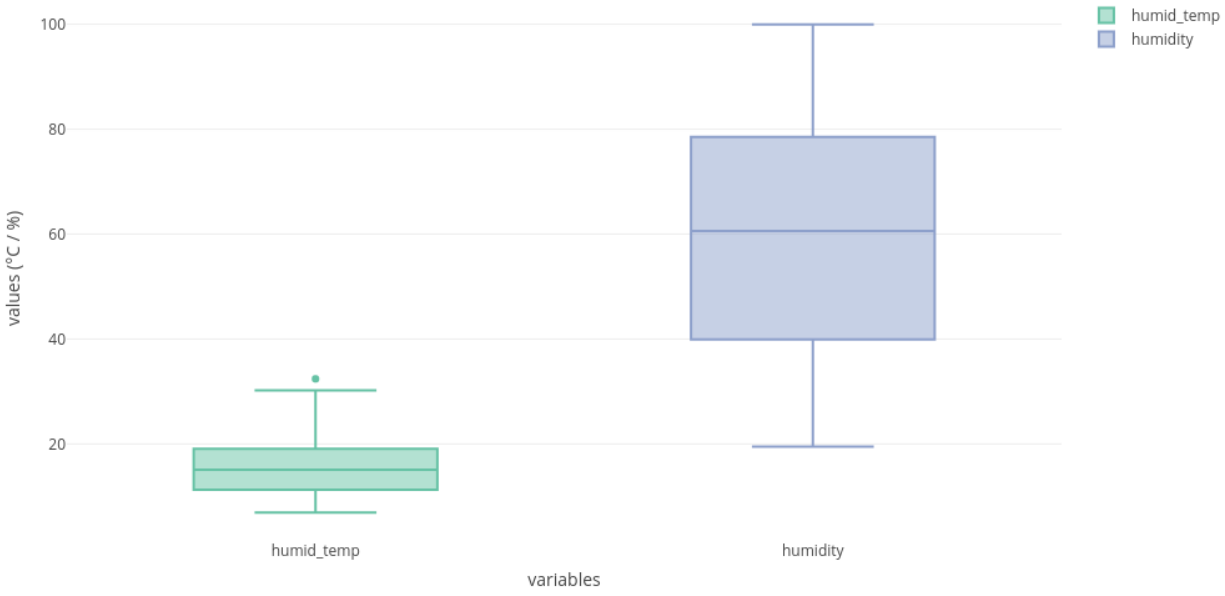
| result_time | epoch | nodeid | humidity | humid_temp | hamatop | hamabot |
|---|---|---|---|---|---|---|
| 2004-05-07 18:34:59.185264 | 2814 | 3 | 97.5805 | 12.5600 | 4070.18 | 0 |
| 2004-05-12 03:15:22.228178 | 4070 | 4 | 75.7940 | 10.4432 | 0.00 | 0 |
| 2004-11-10 14:25:00 | 2814 | 3 | 97.5805 | 12.5600 | 4070.18 | 0 |
| 2004-11-10 14:25:00 | 4070 | 4 | 75.7940 | 10.4432 | 0.00 | 0 |

By inspecting the size of the combined dataset we are given, its row number is exactly the sum of net and log rows. I zoomed in to figure out what's the difference between my merged dataset and the author's way of merging datasets. It turns out that all the duplicates have the same epoch, nodeid, humidity, etc. except the voltage variable. The data point from log dataset has the normal range of voltage while the data point from net dataset has voltage over 200 volts as shown in the example table. I found this issue during the data cleaning process as well, but I didn't generate a new combined dataset. The reason is that the data cleaning procudure automatically removes all the high voltage datapoints. Therefore, data cleaning takes care of both the wrong recording issue and the datasets merging issue.

### 2.3.2 Initial Display of All Data

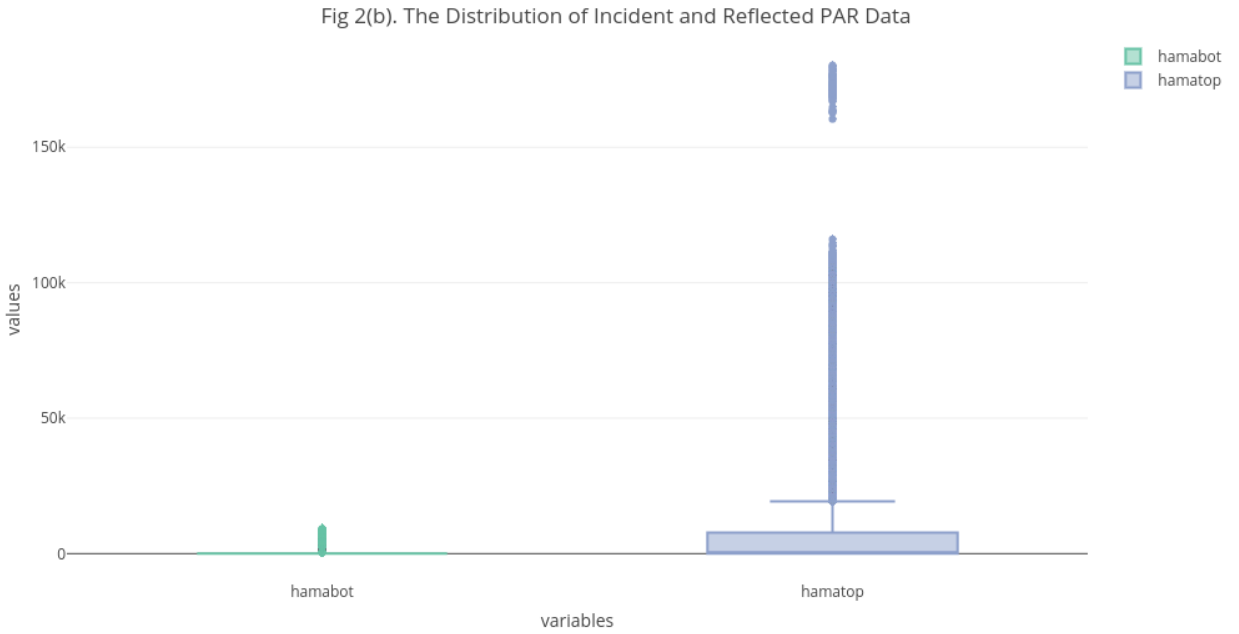
Temperature variable has a median around 15°C, a max around 31°C and a min around 7°C. Humidity variable has median around 60%, max around 99% and min around 19 %. The two variables are normally distributed after data cleaning.

Fig 2(a). The Distribution of Temperature and Humidity Data




**NOTE:**
**This is an interactive plot. Please click the link below.**
**Zoom in one variable by clicking on the legend**

`https://plot.ly/~linqing_wei/1.embed`



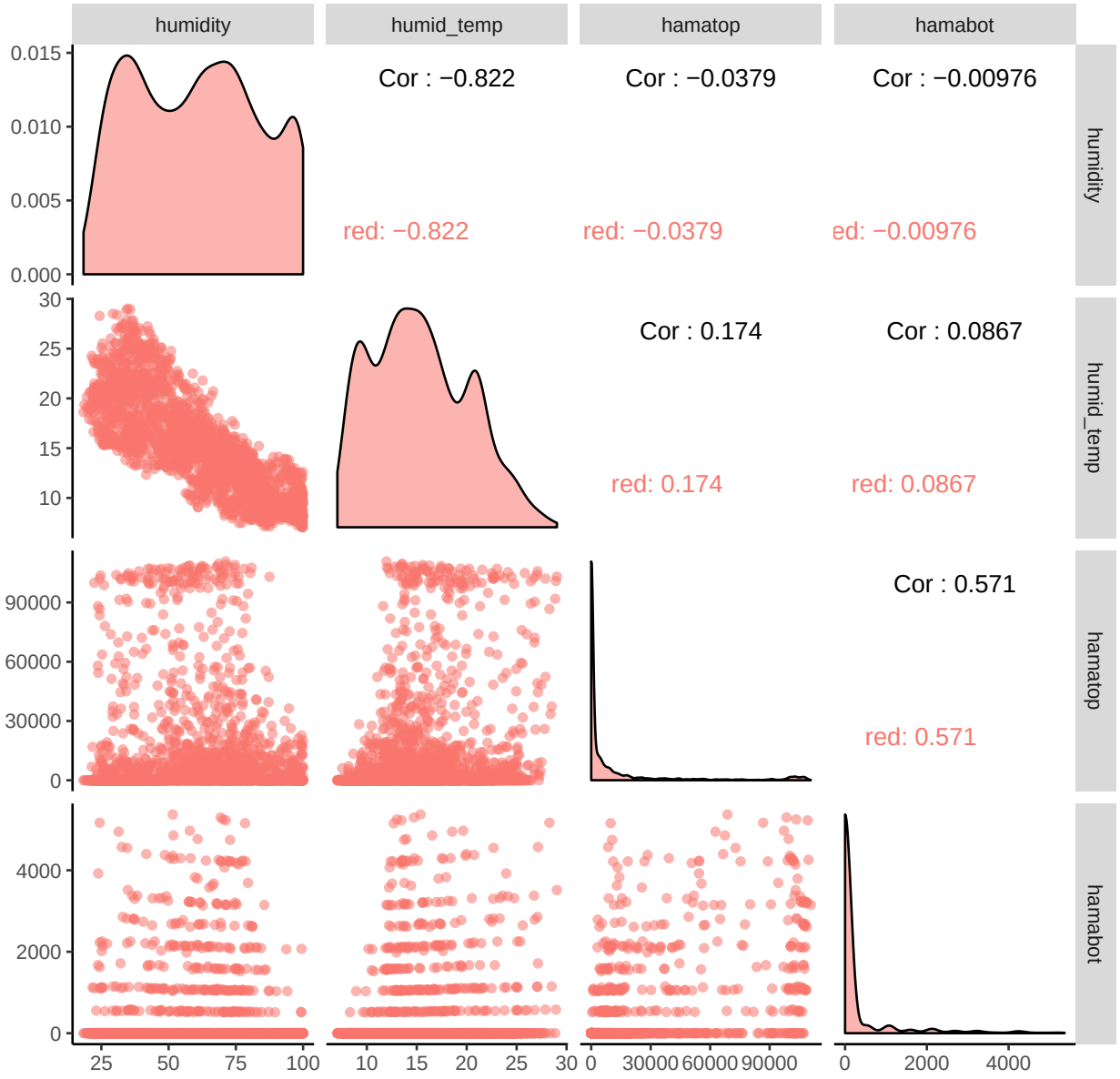

**NOTE:**
**This is an interactive plot. Please click the link below.**
**Zoom in one variable by clicking on the legend**

`https://plot.ly/~linqing_wei/11.embed`

The static image is not a good representation since it compresses the reflected PAR data. Please see the interactive plot and zoom in variable for a better visualization. The plot shows that both incident PAR and reflected PAR have massive amount of outliers forming a straight line. However, since the paper does not provide specific instructions defining the outliers of PAR variables, I would rather keep the outliers rather than remove them. By zooming in each PAR variable, the distribution of both PAR variables are very skewed. The refelcted PAR has a heavy proportion of data centered at 0.

### 2.3.3 Pairwise Analysis: humidity, temperature and PAR

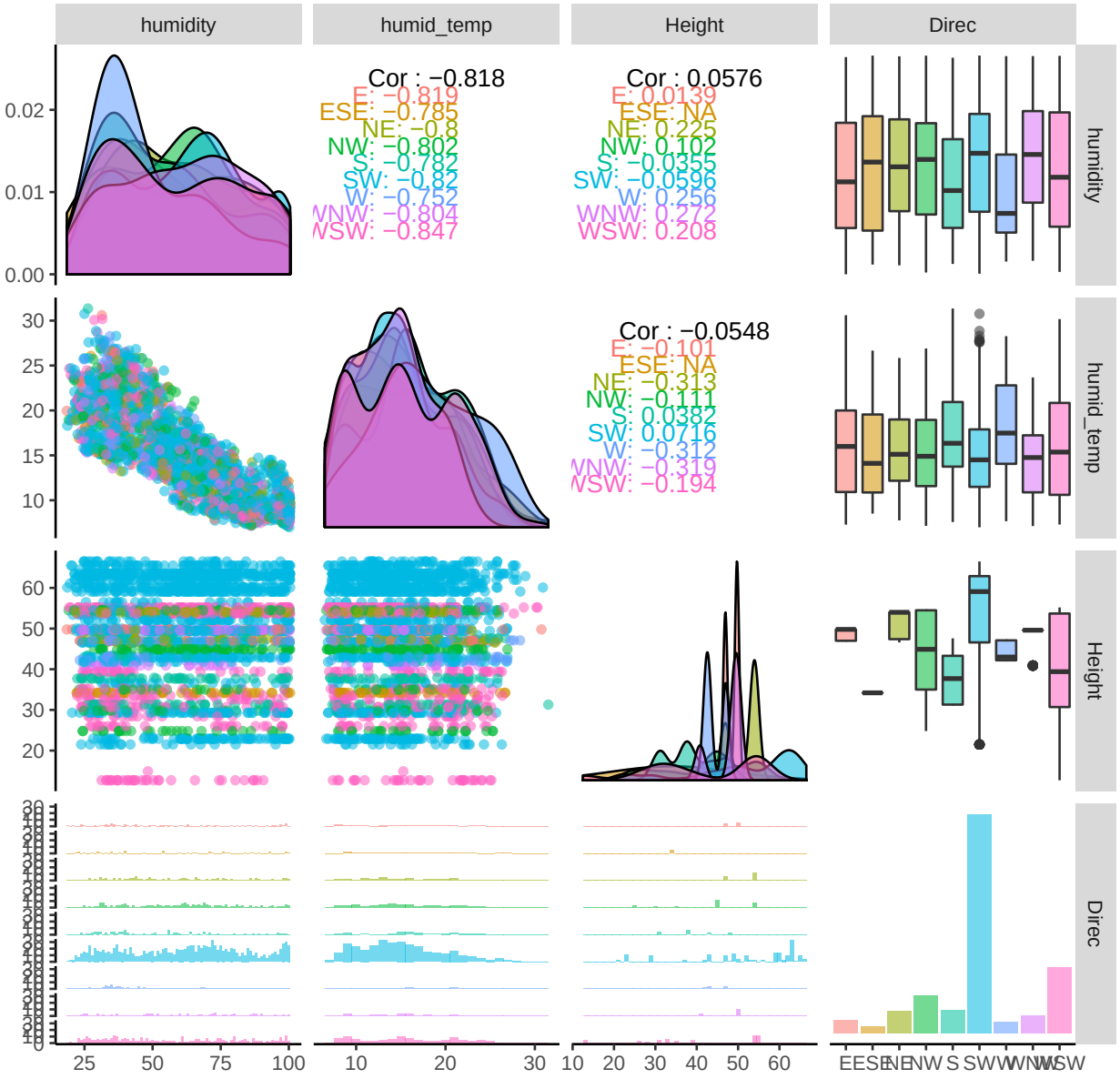
Fig 2(c). Correlations of Humidity, Temperature and Incident/Reflected PAR




Humidity and temperature have a negative correlation while temperature has a positive correlation with PAR. Both incident and reflected PAR's correlations with humidity and temperature are hard to interpret from the plots. It seems that the gradient of humidity and temperature is very patterned, so that we could use this as an anchor point to extend the analysis into temporal and spatial dimensions.

#### 2.3.4 Pairwise Analysis: temperature,Height, Direction

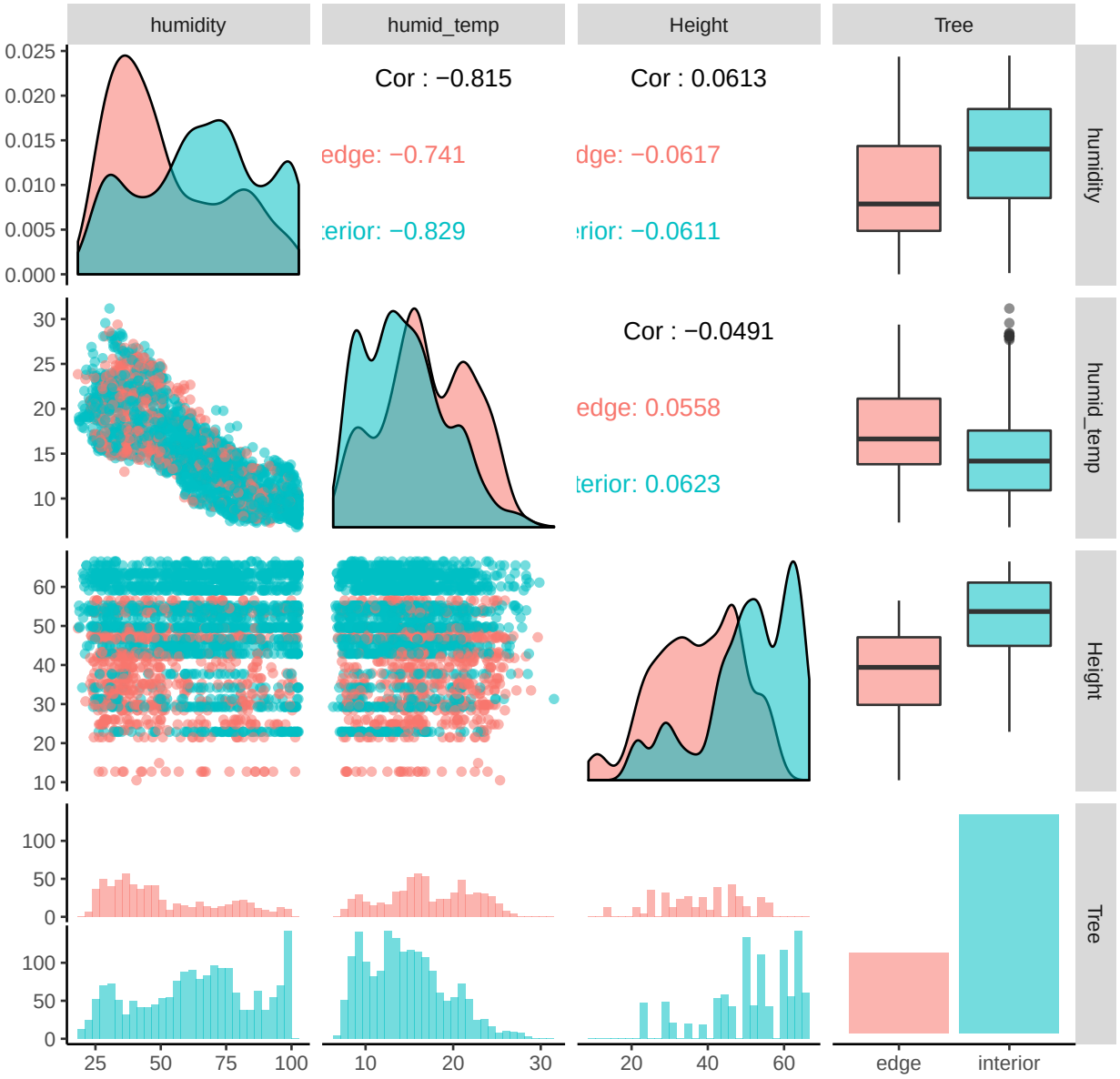
Fig 2(d). Correlations of Temperature, Height and Direction




Humidity and Temperature do have a spatial variation. For example, humidity is lower in the west direction relative to other directions. In the SW direction, humidity data distribution has a larger variance. Temperature distributions in each direction is more consistent. However, in the west direction, overall temperature is higher relative to other directions. The west direction distribution of humidity and temperature encodes the negative correlation shown in the previous section. The spatial analysis on Direction variable could be a future direction to explore.

### 2.3.5 Pairwise: height, temperature, humidity, Tree interior/edge

Fig 2(e). Correlations of Temperature, Height and Tree interior/edge




The interior and edge parts of the tree show very different data distributions. Both are approximately normally distributed, but the interior of the tree has overall higher humidity and lower temperature.

Combining the messages conveyed from data exploration, my focus would be to explore the gradient of temprature and humidity in spatial and temporal dimensions. Also, figure out a way to combine both dimensions to visualize the gradient from different angles. PAR would not be a major focus since intuitively there should be more outlier rejection conducted on PAR data points but there's lack of rejection information from the paper.

## 3 Graphical Critique

I do like the plots in Figure 3 since the representaions migrate from 1D to 2D in a logical order. They try to project the variables on each other to show the internal correlations. However, the same color code in every

single plot blurs out the message that could be conveyed. There's nothing stand out by looking at the plots.

Another downside is the lack of footnotes for PAR. The axis scale of PAR is apprently different from that in the dataset. There's possibly an intermediate unit conversion or data cleaning step to reproduce the plots shown in the paper.

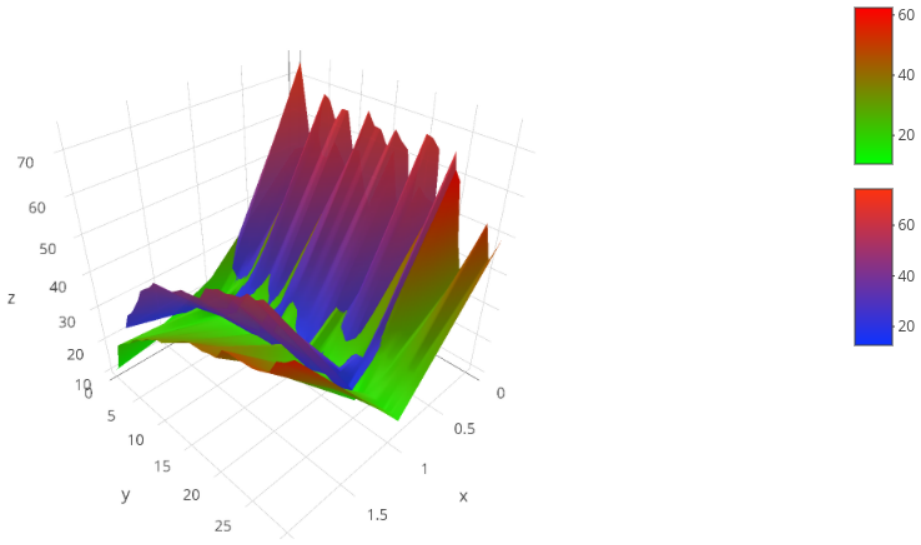
Figure 4 is trying to tell a temporal story for each variable. However, the upper two plots are messy due to the lines. The lower two scatter plots are apprently better to show the time series in a neat way.

# 4 Findings

## 4.1 First finding

I consider spatial analysis as a "scanning machine." The first thing pops up in my mind is to scan the tree from bottom to top, inside to outside, using humidity as the gradient to see the pattern. Tree height and tree parts(interior and edge) are used as the spatial variables. Overall, Tree edge has a lower humidity magnitude than tree interior. Humidity is higher at higher part of the tree, which verifies the positive correlation between humidity and height shown in the data exploration section.

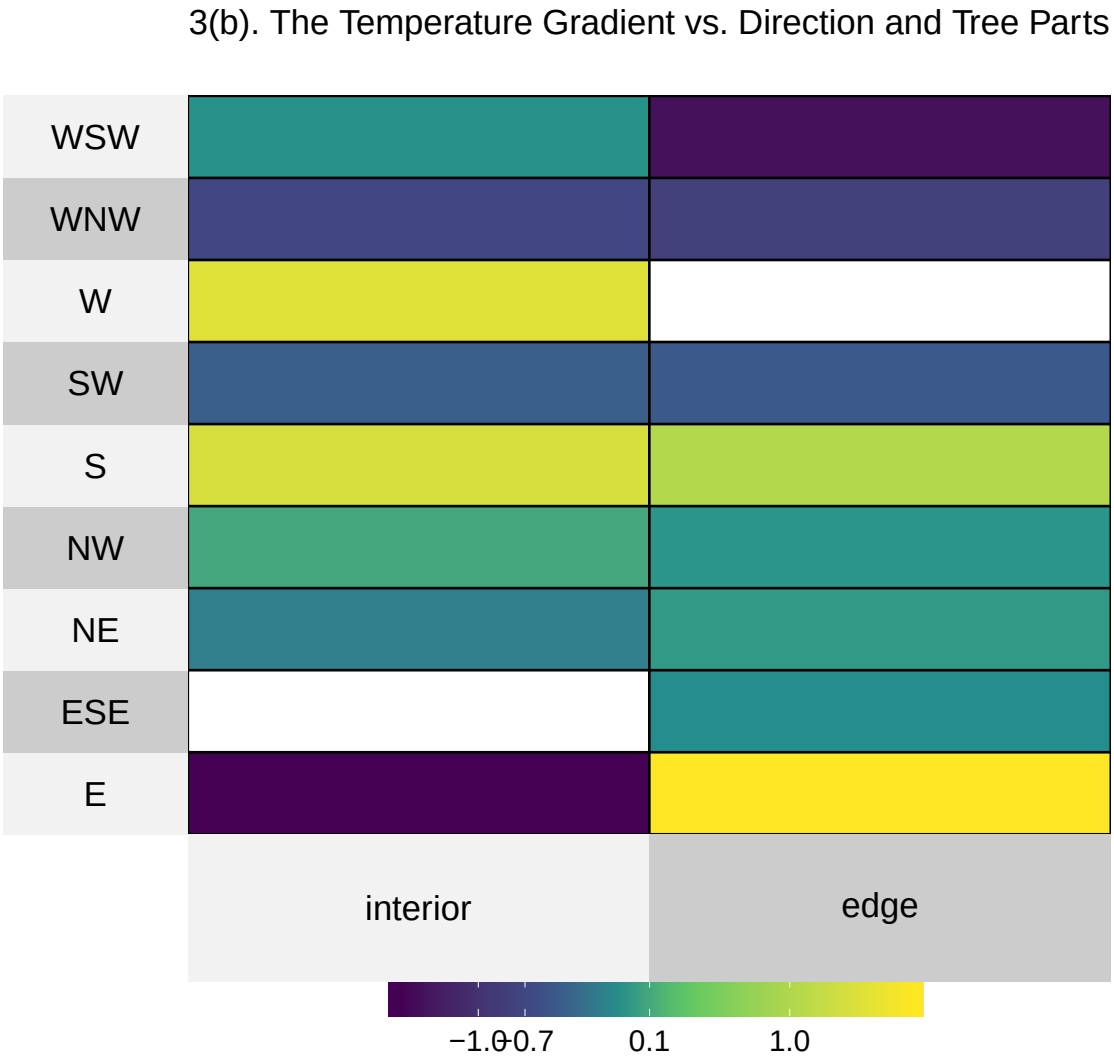
Fig 3(a). Spatial Humidity Gradient vs. Height and Tree Interior/Edge)




**NOTE:**
**This is an interactive plot. Please click the link below.**
**Static graph fails to display the legends.**
**Get a better visualization with interactive plot!**

`https://plot.ly/~linqing_wei/5.embed`

## 4.2 First finding: Part 2



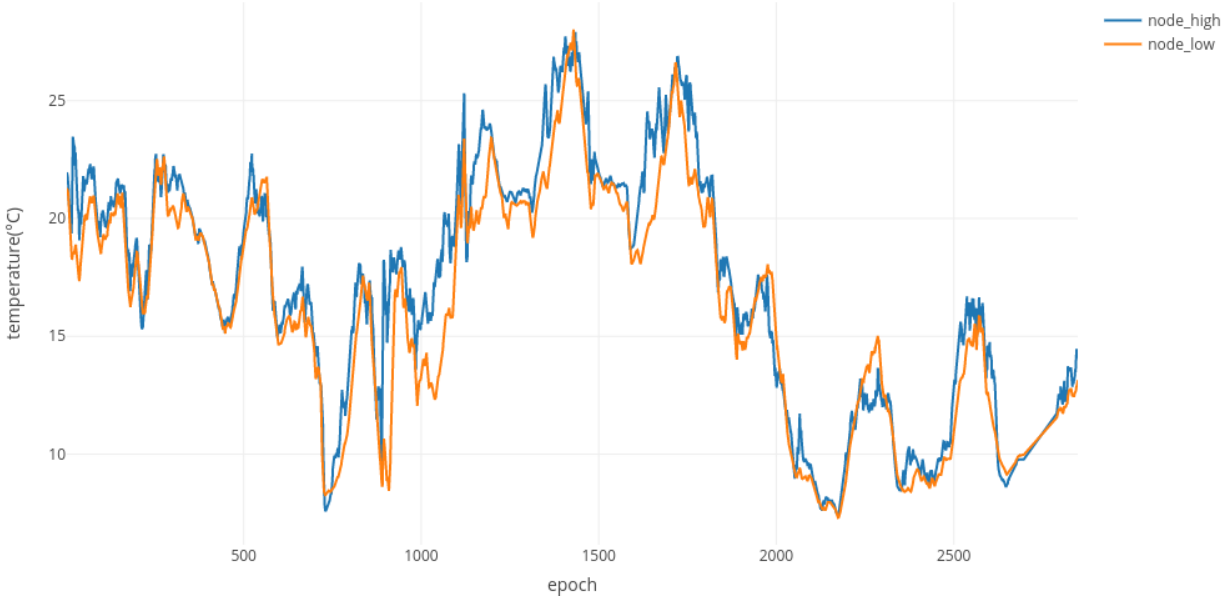

Now we are going to do a 360°C scan of the tree using Direction and Tree parts (interior and edge) as the spatial variables. Intuitively, there should be a temperature gradient in each direction and tree part combination. The heatmap shows scaled temperature intensity, which reveals that the east edge of the tree has the highest temperature. This finding is plausible since normally the east side of the tree gets more solar exposure.

## 4.3 Second finding

The second analysis incorporates both temporal and spatial elements. The motivation of this section is to inspect the temperature change versus time and height of the tree. I chose two nodes, with node 80 placed at the bottom of the tree and node 78 placed at the top of the tree.The node placed at the bottom has an overall higher temperature distribution than the higher node. However, the time series pattern of each node is roughly the same. We could reasonably conclude that the temporal trend is more consistent and stronger than the spatial trend. In other words, the macroclimatic change is more prominent than the microclimatic

change. Although it is a rather obvious fact, it reminds us of the difficulty of conducting this experiment since the macro scale trend could overshadow and influence the judgement of microclimatic trend.

Fig 4. Time Series Plot for Tree Height vs Temperature(°C)



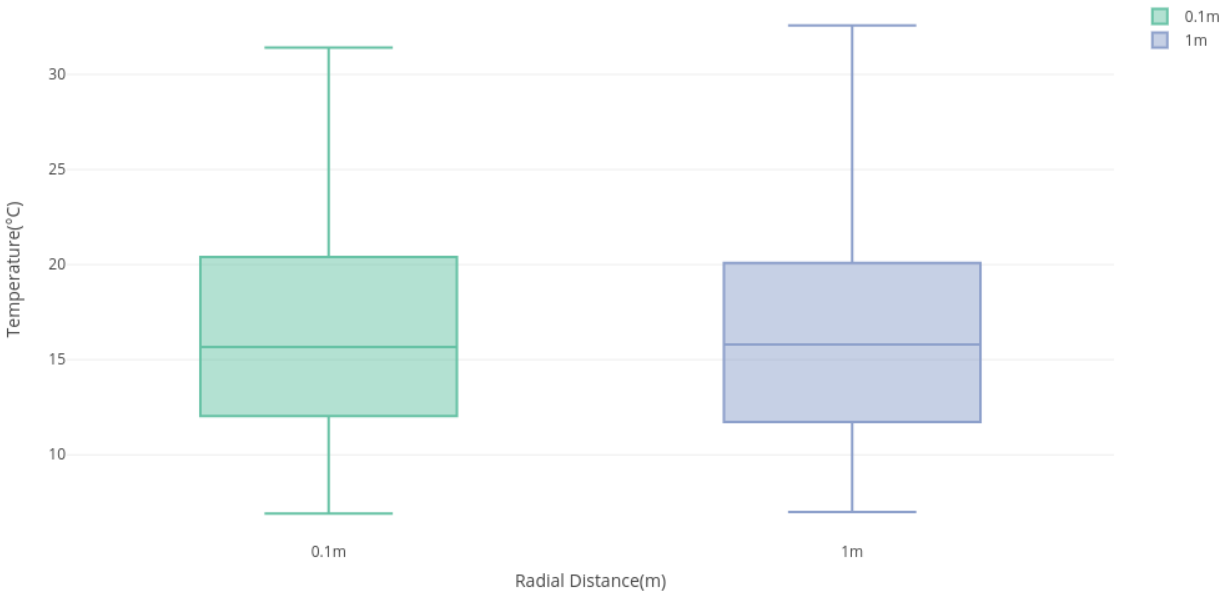

**NOTE:**
**This is an interactive plot. Please click the link below.**
**Get a better visualization with interactive plot!**

`https://plot.ly/~linqing_wei/7.embed`

@

## 4.4 Third finding

Fig 5.The Effect of Radial Distance on the Stability of Data




**NOTE:**
**This is an interactive plot. Please click the link below.**
**Get a better visualization with interactive plot!**

`https://plot.ly/~linqing_wei/9.embed`

This section concerns about the stability and the experimental design mentioned in the paper. Since it is hard to control for the noise in the surroundings, the paper mentioned that they placed the sensors at a very closed radial distance to the tree. I was curious to explore the effect of radial distance on the stability of the data. The plot shows that sensors placed at radial distance 0.1m and those placed at 1m have very similar distributions. However, a slight difference is that sensors placed at 1m does have a larger variability at the tail. Although a definite conclusion regarding the effect of radial distance is hard to reach, it does give us an insight that larger radial distance does affect the quality of data.

# 5 Discussion

The data size is a major probelm when exploring the data. During the exploration phase, the first stage is always to make a very raw plot to include all the data points before downsizing. However, the large dataset takes a fairly long running time. In the later stage, I sampled datapoints from the whole dataset and adjusted the transparency, but with the limited capability of human eyes, the plot's resolution still seems very low. Also, since I decided to create interactive plots to better present the data, I got the second restriction from the intervative plot softwares. There's an enforced max plot size such that I had to keep downsizing the sampling size in order to make compatible plots.

# 6 Conclusion

The redwood data collected from "Macroscope" sensors provide enormous amount of information. However, digging into the underlying relationships requires careful data cleaning procedure. Luckily, the datasets do show several prominent microscopic gradients. Zooming into the spatial dimension, from bottom to top of the tree, humidity increases while temperature decreases. The interior part of the tree has higher humidity than the edge. Also, by segmenting the tree into nine directions, we found that the east edge of the tree has the highest temperature.

The temporal trend is more consistent and identifiable than the spatial trend. It is interesting that when plotting temporal versus spatial trends on the same graph, the deviations in spatial trend appear to be negligible, which in turn shows the delicacy of microclimatic system and the importance of designing more sensitive sensoring technologies.

Taking care of backgroud noise is a major concern of the this experiment setup. Therefore, inspecting the data stability gives us some intuition regarding the experiment setup. The sensors are placed at a radial distance very close to the tree, but it does slightly affect the data variability.

The future analysis could move onto two tracks: experimental design and data analysis. For the experimental deisgn part, I noticed that the authors mentioned the difficulty of zooming into the microclimatic system. The experiment is lack of a rigorous control group to eliminate the background noise, which makes it hard to distinuigsh between local trend and external macrosopic trend.

In terms of future data analysis, I would put a heavier emphasis on the temporal analysis. The current analysis does not dig into the dates dataset too much. Several interesting questions we could ask are: 1. How does the temporal pattern change during summer solstice? 2. Would the trend differ if we plot time series versus Direction? 3. Pick a day with the most dynamic climatic change. How fast or sensitive does the redwood microclimatic system adapt to the macroscopic change?

This research does provide us a great insight on the microclimatic system of redwood trees, and the rich amount of data obtained from the sensor network is a valuable source of revealing the hidden patterns that

we could not directly interpret via human eyes.

# References

[1] Gilman Tolle et al. *A Macroscope in the Redwoods*. Proceedings of the 3rd International Conference on Embedded Network Sensor Systems, San Diego, CA, 2005.