STA 521: Project 1 Redwood Data Report

Xiaozhu Zhang (2625755) and Xuyang Tian (2328958)

September 2021

1. Data collection

1.1 Summary about the paper

This paper reports on a case study of microclimatic monitoring of a coastal redwood canopy. Since the redwood trees are really tall, usually 50 meters to 70 meters, there is substantial variation and substantial temporal dynamics in a single entire redwood. Today, thanks to the development of "macroscope" sensor, we have obtained more accurate as well as detailed data and thus an unprecedented picture of environmental dynamics over such a large organism by using a large number of wireless micro-scale weather stations.

The experiment in this paper is carried out in a study area in Sonoma California, over a 44-day time period. Our main data is: a month in the life of a redwood tree. We deploy 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. Each node measured air temperature, relative humidity, and photosynthetically active solar radiation.

The top of the tree experiences wide variation in temperature, humidity, and, of course, light, whereas the bottom is typically cool, moist, and shaded. This variation was understood to create non-uniform gradients, essentially weather fronts, that move through the structure of the tree. For example, as the sun rises, the top of the canopy warms quickly. This warm front moves down the tree over time until the entire structure stabilizes or until cooling at the canopy surface causes the process to reverse.

Through this experiment, scientists have verified the existence of spatial gradients in the microclimate around a redwood tree. Having captured enough data to track the changes in these gradients over time, we can begin using this data to validate biological theories. For example, our newly-obtained data from the macroscope could be used to build a quantitative model of the effect of microclimatic gradients on the sap flow rate. With a more detailed understanding of sap flow and transpiration, biologists can also work toward understanding the large-scale processes of carbon and water exchange within a forest ecosystem.

1.2 Summary about data collection

In a nutshell, the data are collected from a wireless sensor network that recorded 44 days (Apr 27 to Jun 10) 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. Our main variables of interest are air temperature, relative humidity, and photosynthetically active solar radiation, which mostly effect the daily life of a redwood tree.

The major difference between sonoma-data-net.csv and sonoma-data-log.csv is that the former one are data retrieved over the wireless network while the latter one are data retrieved from the flash logs after the deployment. These two tables comprise the sonoma-data-all.csv together. The deployment details are as following:

1) Time: One month during the early summer, sampling all sensors once every 5 minutes. **2) Vertical Distance:** 15m from ground level to 70m from ground level, with roughly a 2-meter spacing between nodes. **3)** Angular Location: The west side of the tree. **4)** Radial Distance: 0.1-1.0m from the trunk. Prior to deployment, we elaborate on a comprehensive deployment strategy that carefully tests and calibrates the sensors. Our strategy used two calibration phases: roof and chamber, which target at different variables of

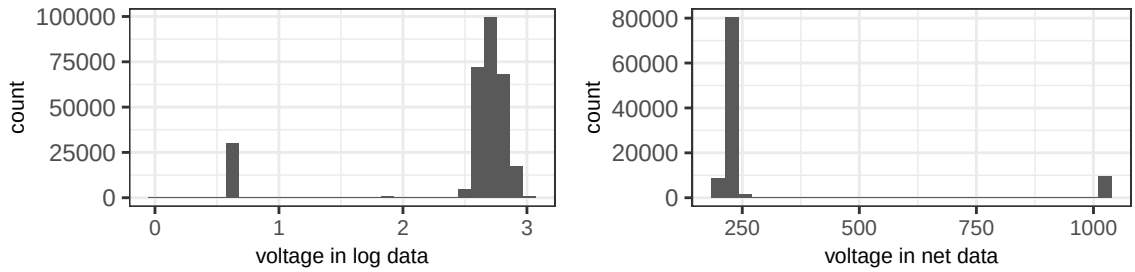
interest. Each phase provided performance data on different subsets of the sensors installed on the mote. The roof calibration provided a real-world data source for the PAR sensors – direct sunlight; while the purpose of the chamber calibration phase was to understand the response of the temperature and humidity sensors to a wide range of phenomena.

2. Data cleaning

2.1 Histograms checking

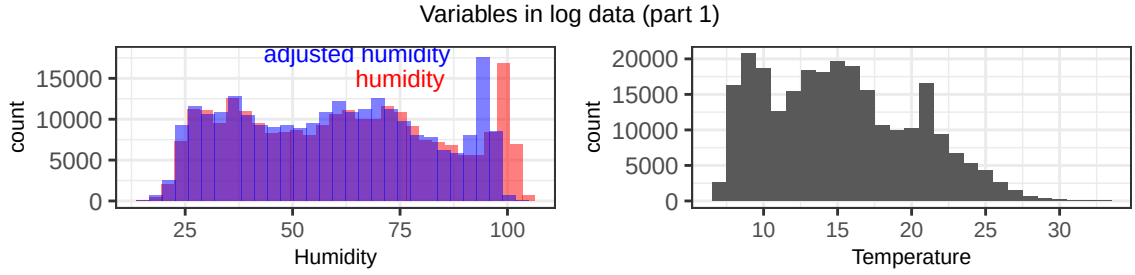
In this section, we are going to do some basic histogram checks as the start of EDA. However, before drawing any plots it is necessary to look into the variables provided in the two files at first, especially in the case when the data dictionary is not available. Among the 11 variables, `epoch` determines the time point when the data is collected, while `nodeid` determines the spatial point where the data is collected. Notably, `epoch` and `nodeid` combined provides a unique identifier for one observation. Interesting enough, the `result_time` in the `sonoma-data-log` file are not the time when data were collected, but might be the time when data were downloaded from the logging backup; therefore we must re-match the date and time in `sonoma-data-log` (and also `sonoma-data-all`) according to the `epoch` variable and the `sonoma-dates` file. The `parent` and `depth` variables corresponds to the "tree structure" of network deployment and signal transmission, while do not provide useful information for our task here.

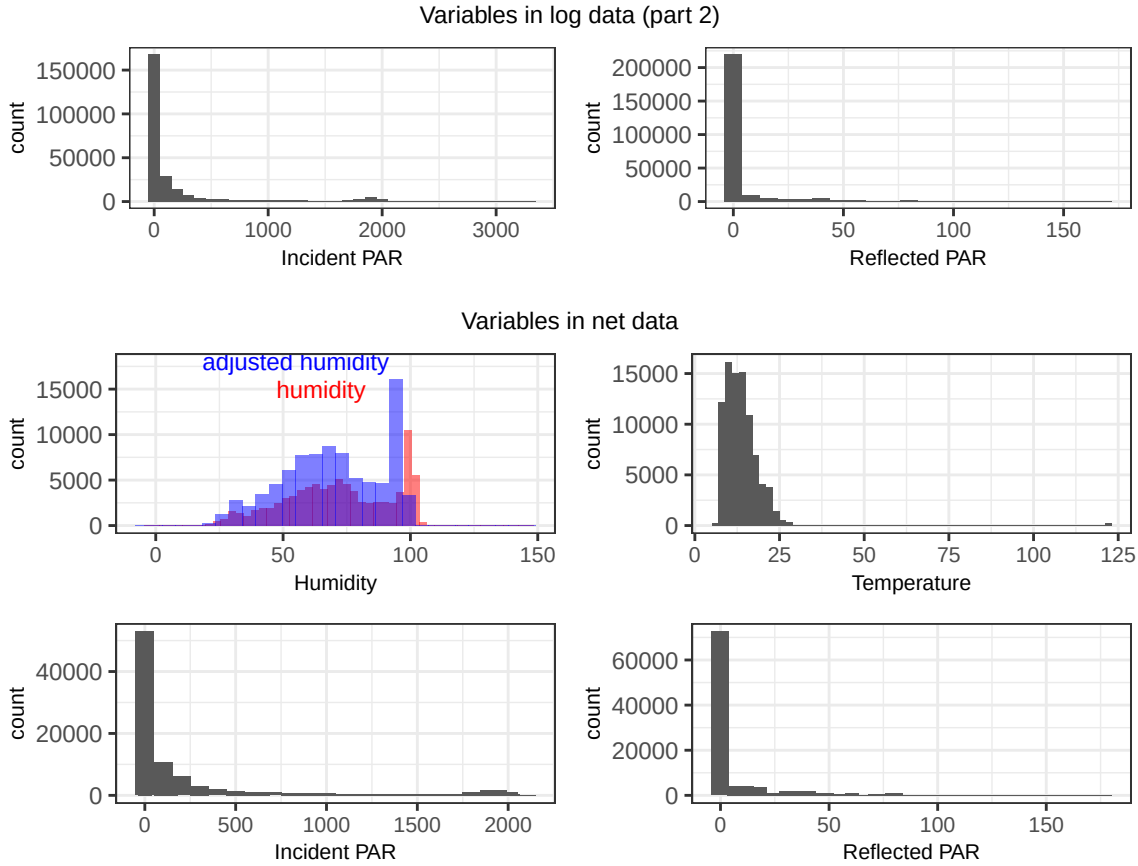
As for the variables we are going to plot out, we first check the `voltage` from the files `sonoma-data-log` and `sonoma-data-net`.



It is notable that: (i) both `voltage`s have outliers; (ii) the `voltage` in the two files are of different scales. According to Tolle (2005), the reasonable `voltage` should be within 2.4V to 3V. Our strategy here is to remove all observations outside this range in `sonoma-data-log`, while only remove visible right-tail outliers in `sonoma-data-net` since the relationship between the two scales is unknown for now. The conversion and further analysis will be discussed in part (e).

Using the data whose voltage-related outliers are removed, now we can plot out the histograms of remaining variables. In order to keep the unit of PAR consistent with that of Tolle (2005), we use the formula $1\ \mu mol\ m^{-2}s^{-1} = 54\ Lux$, where the light source is sunlight, to convert the unit of Lux in the raw data to the unit of $\mu mol\ m^{-2}s^{-1}$.





2.2 Removal of missing data

Note that there are many repeated records in both files, so before the beginning of the whole analysis we have already removed the replicates. After the removal of both replicates and outliers, now there are 6898 (2.63%) observations with missing values in the `sonoma-data-log` file, and there are 3710 (4.08%) observations with missing values in the `sonoma-data-net` file. Due to the rather small number of missing values, the removal of these records shall not influence the data overall effectively.

The plots above showed the distribution of missing values in each of the two files. Mostly, the missing values

were evenly distributed among the full range of hours, days and dates; except for the 6th day and 26th–30th day in log data, and the 16th day and the 28th–7st day in the net data. These are the days when nodes were less saturated and logging backup worked well.

2.3 Incorporating location data

There are at least 82 nodes that were placed in the network deployment. However, records from only 80 nodes were collected, among them information of two nodes (ID = 100, ID = 135) were missed in the file `mote-location-data`. After incorporating the location data into the main table, there are 15 variables in total in the new data frame.

2.4 Easy outliers identification

Based on the histograms in section 2.1, we can eyeball several easy outliers that are far away from the main cluster. The variable humidity has two versions, where the adjusted one is clearly shifted to the left from the original one. Based on the "calibration analysis" from Tolle (2005), along with the fact that relative humidity should be in the range of 0%–100%, we should discard the original humidity and use the adjusted one instead. In addition, any value of the adjusted humidity below 0% and above 100% should also be removed.

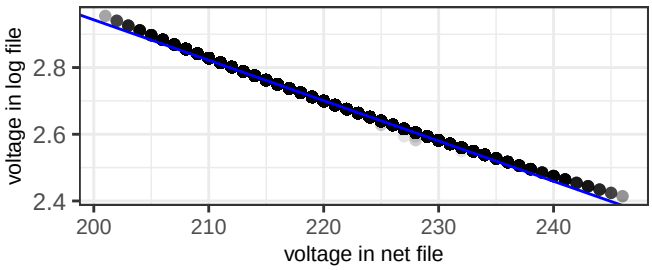
As for the temperature variable, in the file of `sonoma-data-net`, 75% of the data are below 16 Degrees Celsius while the max is around 122. Hence we identify this unreasonably large data point as an outlier, and remove all records with temperature above 40 to adhere to the common sense.

For the variable incident PAR, we recognize it as an extremely right-skewed distribution. Around 50% of the data points are 0, while the max in the two files are around 3330 and 2100 respectively. We zoom in the histograms using `coord_cartesian()` in `R` and found that there are no values between 2250 to 3000 in the Incident PAR plot for log data, indicating that the clusters around 3000 are outliers for sure. However, in order to find a reasonable cutpoint we have to have some expertise knowledge. We decide that any value above 2250 is an outlier by refering to the Tolle (2005).

For the reflected PAR, we use the same strategy as that for the incident PAR. Interestingly, when we zoom in the histograms, we found that there are continuous values between 50 to 150 with non-zeron frequency. Similarly, using the prior information from Tolle (2005), we regard any value above 180 to be an outlier.

2.5 Other outliers identification

In this section we are going to look for more possible outliers based on the variable `voltage`. Recall that the `voltage` variable has two different scales in the file `sonoma-data-log` (the magnitude of 2) and `sonoma-data-net` (the magnitude of 200), while the reasonable voltage of node battery should be in the range of 2.4V to 3V. In order to identify the relationship between the two scales, we first pick out the records that exist in both files, using the identifier (`epoch`, `nodeid`). Notably, these records have the same variable values in both files except for the variable `voltage`. With such a discovery, we can plot out the scatter plot of the two scales, based on which an almost perfect line, `voltage` in log file = 5.36305417 - 0.01209994 `voltage` in net file, is fitted.



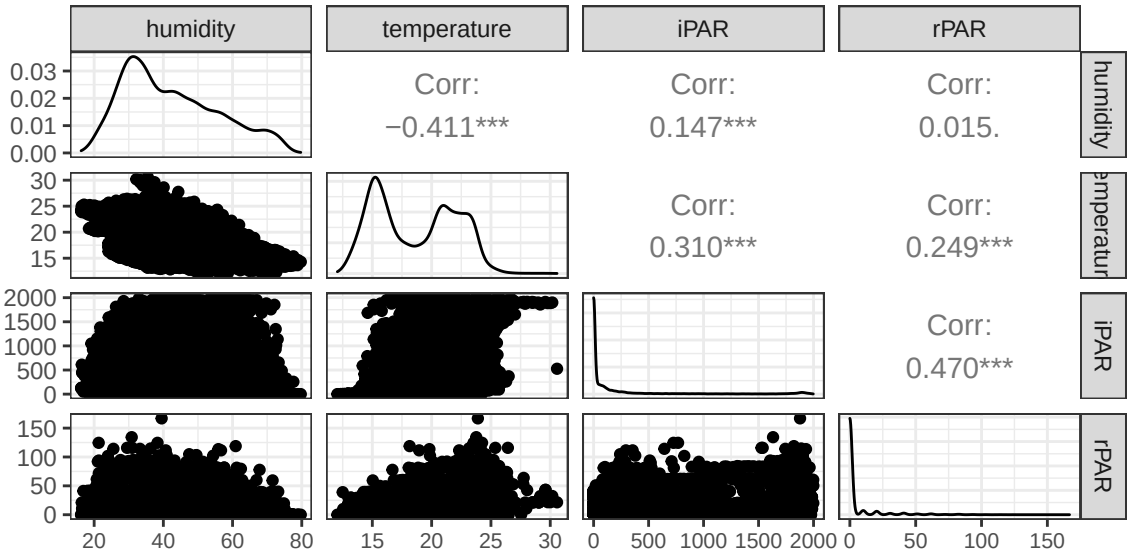
Using this relationship in hand, we can convert those `voltage` in the net file from the magnitude of 200 to the magnitude of 2, and then combine the two files into a single clean data frame. Repeated records identified

by the same (`epoch`, `nodeid`) must be removed. In this new data frame, finally, the records with `voltage` out of range (from 2.4V to 3V) can be again recognized as outliers and shall be removed.

3. Data Exploration

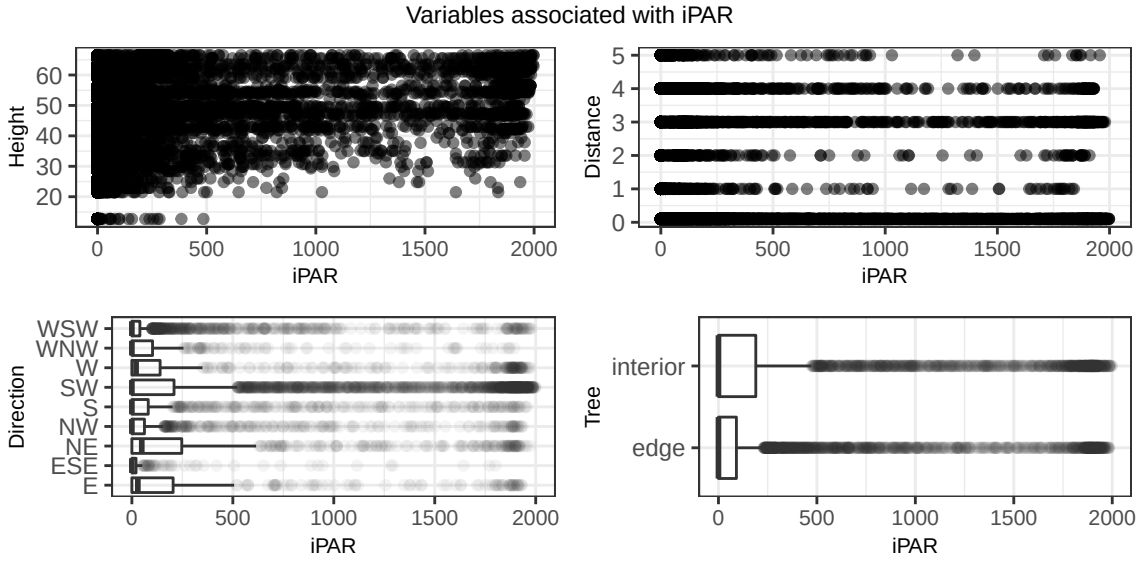
3.1 Pairwise Scatterplots

In this section, we are going to make pairwise scatter plots of variable humidity, temperature, iPAR and rPAR. Among the approximate 274 thousand records across 44 days, we select the day May 1st using the same reason as Tolle (2005) – "it contains a wide range of temperature and humidity readings throughout the course of the day". The distribution density of each variable, scatter plots and correlation coefficients are shown clearly in the plots. It is reasonable to find that temperature and humidity are strongly and negatively correlated, and the two types of PAR are strongly and positively correlated. In addition, it is quite interesting to find that, temperature is mildly and positively related to the two types of PAR, while humidity is more or less weakly but still significantly correlated to iPAR. It is not hard to explain these findings from a biology perspective, since below the threshold of 40 degree the rate photosynthesis generally increases as the temperature increases. Due to the fact that less water from the plant would evaporate when the air around the plants has lots of humidity, and that water is one of the raw materials of photosynthesis, the rate photosynthesis generally would increase as well as the humidity increases.



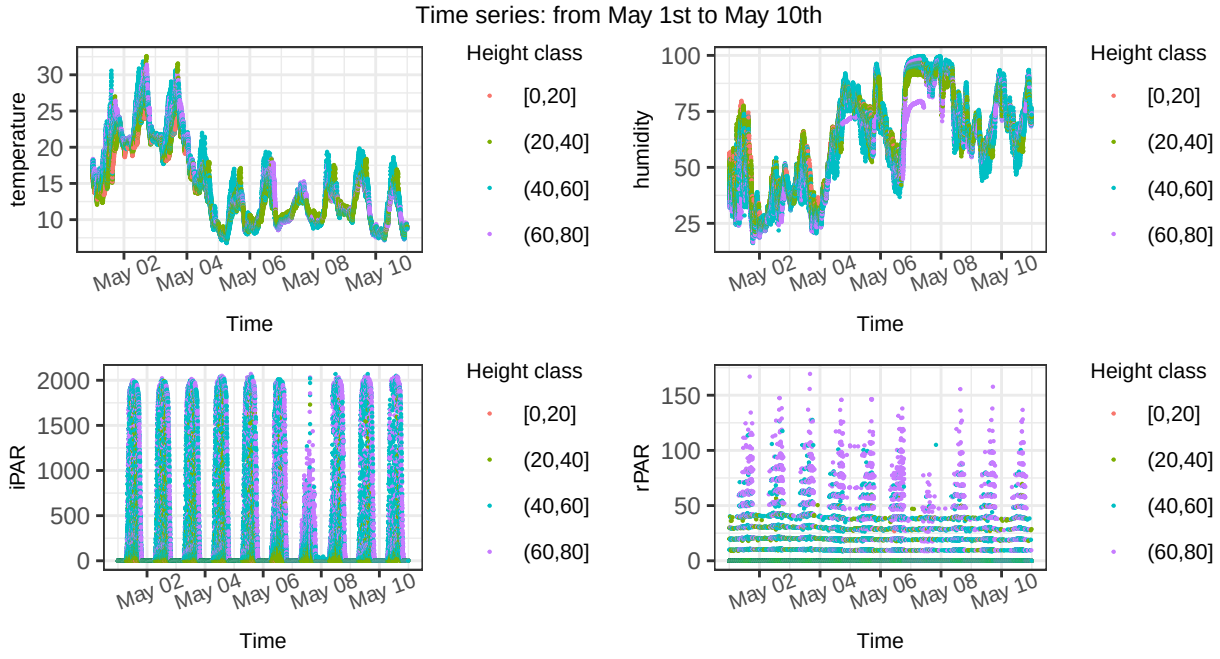
3.2 More variables associated with iPAR

In this section we shall find more variables associated with iPAR besides the variable humidity, temperature and rPAR that we found in the previous section. Much information can be revealed by the graphs below. The upper left scatter plot reveals that iPAR and height have a clear positive correlation. Specifically, the higher the node is, the wider the range of iPAR would be, indicating a higher chance of detecting high iPAR. From the two lower boxplots, we can tell that the directions northeast, east, and particularly southwest generally receive more iPAR, while interestingly the interior of trees receives more iPAR than the edge of trees. However, no particular pattern between iPAR and distance can be recognized from the upper right scatter plot. In section 4.2, we will dive a little bit deeper regarding the associations, and try to form reasonable explanations.



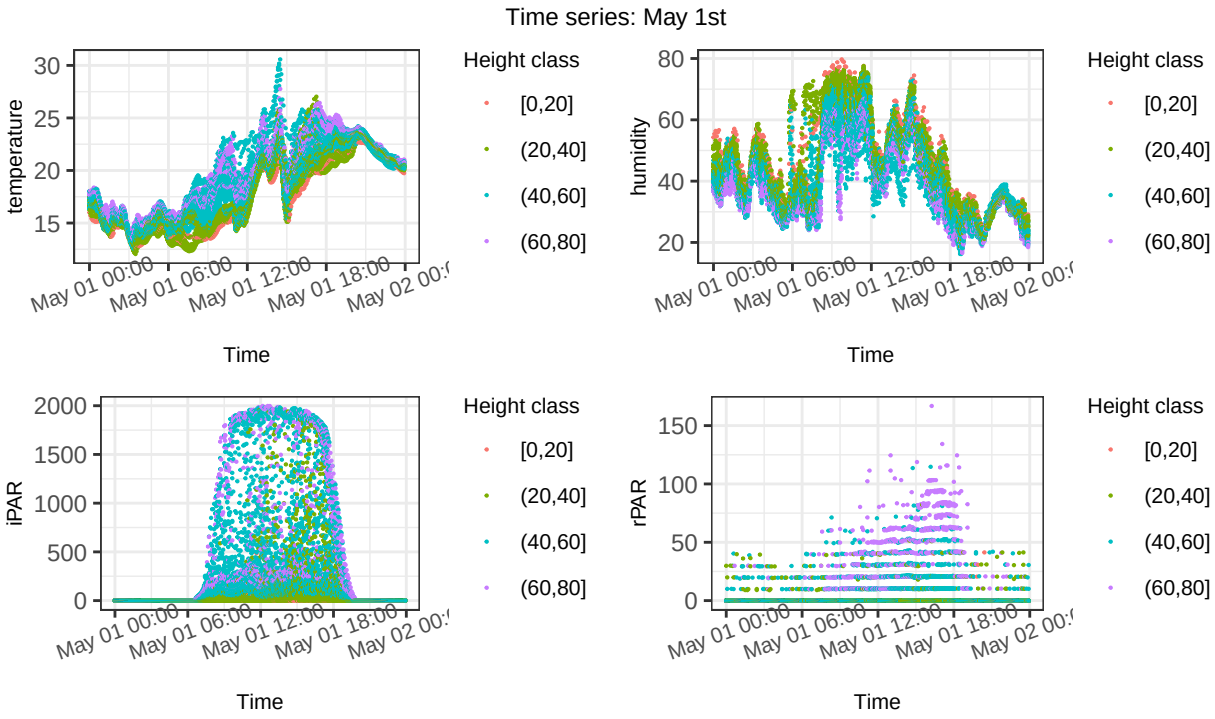
3.3 Temporal and spatial trends of variables

In this part, we are going to look into the temporal and spatial trends of the four interested variables. The plots of time series are plotted with the levels of height as color cue for temperature, humidity, iPAR and rPAR. We choose to do it in three time scales: (i) 10 days from May 1st to May 10st; (ii) 1 day – May 1st; (iii) 1 hour from May 1st 12:00 to May 1st 12:55. The three scales provide us with different scopes to explore any underlying behaviors of the variables.

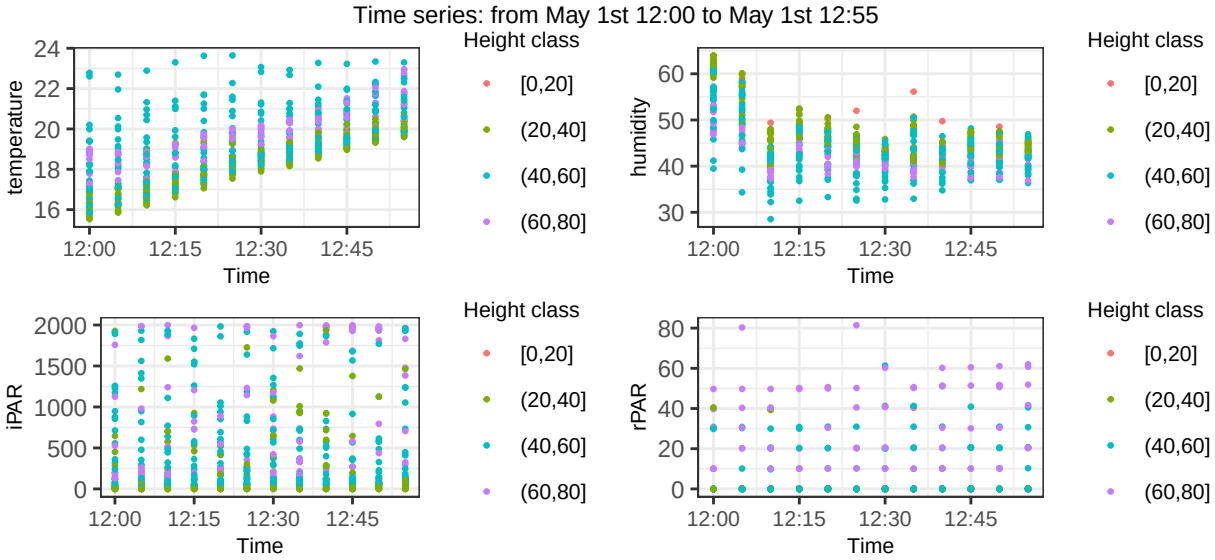


The plots for 10 days give us a macro-scope of how the interested variables changed generally. Although colors are stacked with each other so do not provide much information, we still found that, temperature, iPAR and rPAR periodically increased and decreased on a daily basis, which aptly corresponds to the rising and setting of the sun. In the 7th day both iPAR and rPAR reached to a lower peak, while the humidity reached the highest peak, indicating a rainy weather. In addition, across the 10 days the temperature has a

clear decreasing trend, while the humidity enjoys an increasing trend. The common sense is verified that they are negatively correlated.



The plots for 1 day provides a mild scope. We observed similar graphs as Figure 4 in Tolle (2005). Nodes placed in higher places detected higher temperature, while detected lower humidity across the day. In the morning, most high iPAR records were detected by nodes at least 40 meters high, while in the afternoon nodes above 20 meters high also detected strong iPAR. However, for rPAR, although higher nodes detected higher rPAR in the day, it was not the case at night. Before sun rising and after sun setting, rPAR was similar among all height levels.

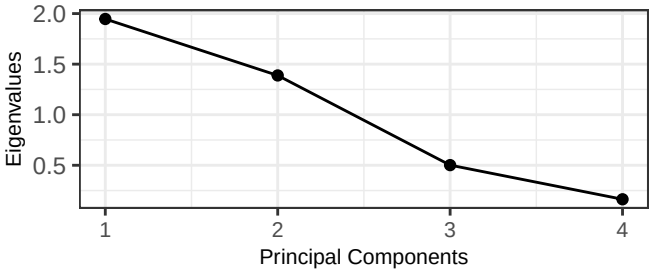


The one-hour plot gives an extreme micro-scope, which leads to many interesting findings. At noon the temperature is gradually increasing, which makes sense, but nodes at 40-60 meters high detected significantly

higher temperature than nodes at 60 meters above did. It is also strange that, the nodes at 40-60 meters high detected much lower humidity than nodes at 60 meters above did. We shall elaborate these two findings in section 4.1. The range for iPAR and rPAR did not change much, because the whole hour at noon is the active time for photosynthesis. Generally speaking only nodes above 20 meters high could detect iPAR and rPAR, and the quantity is positively correlated with height levels.

3.4 PCA

In this section we conducted PCA using the four variables: temperature, humidity, iPAR and rPAR. According to the scree plot below, we can observe a clear decreasing trend of the eigenvalues, indicating that the first two or three principal components are able to represent all information. If we look into the loading matrix, the PC1 mostly incorporates information of humidity and temperature, while PC2 mostly includes information of iPAR and rPAR. PC3 is similar to PC2 and PC4 is similar to PC1. From this perspective, we believe that the first two principal components are enough, and can be used to construct a two-dimensional basis for the whole data.



4. Interesting Findings

4.1 Finding 1

We previously hypothesized, based on the common sense, that the redwood's top part, due to more explosion to sunlight, would always be the most humid and hottest among all four height classes throughout the day. However, the graphs in section 3.3 suggest a different conclusion. When it's around noon, the nodes placed in the 40-60m height class rather than the 60m-above height class detected the highest temperature and lowest humidity instead. Even though this finding sounds counter-intuitive at the first glance, it actually implies the plants' wisdom to protect themselves from overheat – transpiration. The leaves at the canopy experience far more intense transpiration to balance out the direct sunlight and thus have a relatively lower temperature because of the transpiration's cooling effect. In addition, water evaporated surrounding the chunk leads to the higher humidity we observed. On the contrary, the 40-60m interval experiences a far more gentle transpiration and then keeps hot and dry.

4.2 Finding 2

In this part we would like to explore more on the factors associating with PAR, as an extension of section 3.2. Note that in the previous part almost all conclusions were made based on visualization, while no rigorous quantitative tests were involved. In this part, we carried out ANOVA of iPAR against `Dist` (distance of nodes from tree), `Dire` (direction of nodes), `Tree` (position of nodes including edge and interior), and their interaction terms. The result shows that all p-values are close to 0.

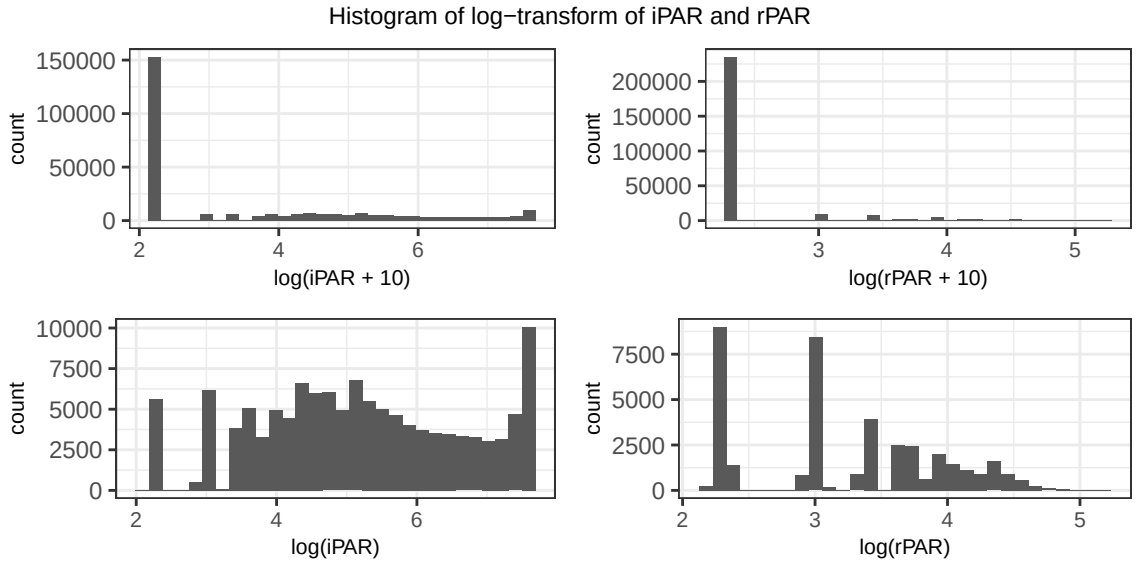
We can conclude that the redwood's photosynthesis is quite three-dimensional in space: it is not only influenced by the leaves' relative distance and position to the trunk but also the direction of sunlight they face. As our coefficient for position suggests, leaves at the interior of the tree experience a much stronger photosynthesis on average than those at the edge: 229 vs 131. The distance coefficients reveal an interesting story: as the distance from the trunk increases, the photosynthesis first intensifies and then weakens: corresponding coefficients are 206, 59, 185, 296, 179, 141. A possible explanation is that the closest and furthest ends are all

covered by the tree itself or other trees, leaving only the middle part unshaded. For the direction parameter, we surprisingly found that the strongest photosynthesis take place on leaves facing NE (324) and SW (231). These two directions correspond to the sunrise and sunset time respectively. We propose that the redwood slows down its process of gaining nutrition from sunlight at noon to avoid overheat and then resumes when it's turning cooler to acquire enough glucose. The interaction terms tell a similar story: `Dist` (level 3) × `Dire` (level S) has a coefficient of 292, and `Dist` (level 3) × `Dire` (level W) has a coefficient of 237.

5. Graph Critique in the paper

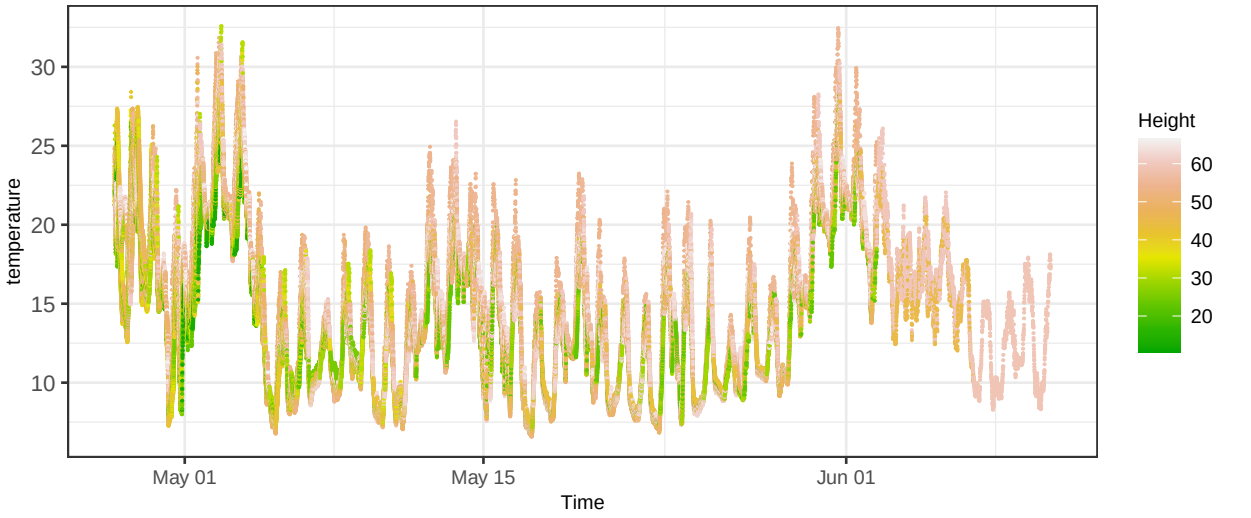
5.1 Log-transform of data

In this part we are going to make a better plot to visualize the distribution of iPAR and rPAR. By "better" we mean full information can be read from the plots. Note that iPAR and rPAR can take value of 0, so first we choose the transform formula $\tilde{x} = \log(x + a)$, where $a > 0$. However, with too many zeros the resulting plots are still right-skewed and hard to read. Hence we decide to make two more similar plots where all zeros are removed, as a zoom-in version of the upper two plots for the sake of clear reading.



5.2 Critique of 2D plots

For 3(c) and 3(d), the author aims to project the four main variables of interest on a 2D plan by squeezing the 3rd dimension, height of the sensor, in the form of boxplots. The difference between 3(c) and 3(d) is that 3(c) displays the original data while 3(d) trying to show us the constant trends along the height by centralizing the data. However, ignoring the time dimension may be misleading since the relationship between height and each variable can be different in different periods of time. Inspired by Figure 4 in Tolle (2005), we incorporated color into a time series plot so all 3 dimensions are included. According to this plot, we found that at night lower part of trees tends to have higher temperature, while during the day higher part of trees are warmer.



5.3 Suggestions for improvement of Figure 4

For the upper left two plots, each line represents the readings taken by an individual sensor. However, we do not know how they correspond to a specific height value. In other words, the information about heights is missing and thus we cannot see any trending related to height from these graphs. In addition, the lines at the beginning and end are intertwined and highly overlapped, so we cannot see the colors very clearly.

For the upper right two plots, the main focus should be the regression line rather than the points, but green is too faint to see. Thus, we suggest making the line much thicker and in a bolder color. For instance, we can make the line red while points black.

5.4 Critique of Figure 7

Figure 7 mainly shows how the yield data varies over time and space, and our main focus is on highlighting the difference between the performance by network data and that by log data.

For the first graph, we can replace the histogram plots with density plots and put them together. Then we can see the bimodal properties of these data better as well as the shift between these two sources of data. For the second graph, we think that the difference is already obvious enough. We can clearly see that most sensors "die" on May 26 from figure 7(b).

In the third graph, we can put all these points into a single graph while they are labeled by different colors. Thus, we will get a better understanding of how their ways of clustering differ. In the fourth graph, we can also make the bars showing network data and log data together side by side for each height. Thus, we will get a more distinguished observation.

Reference

1. Tolle, G., Polastre, J., Szewczyk, R., Culler, D., Turner, N., Tu, K., Burgess, S., Dawson, T., Buonadonna, P., Gay, D. and Hong, W., 2005, November. A macroscope in the redwoods. In *Proceedings of the 3rd international conference on Embedded networked sensor systems* (pp. 51-63).