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

September 16, 2021

## 1 Introduction

In Tolle et al, the authors describe how deploying a wireless sensor network onto a redwood tree can aid in being able to accurately describe the dynamics associated with the microclimate (air temperature, relative humidity, and photosynthetically active solar radiation) surrounding the tree. Recall that redwood trees are extremely tall and relatively wide. The use of wireless sensors enables fine-grained collection of climate data along and near the trees without risking human injury or damage to the tree. The ability to gain a wealth of data in this space can help better understand how climate change can affect redwood trees. Expanding this technology to capture a wider number of trees in a redwood forest would also yield broader data on how a redwood forest behaves.

## 2 Data

In this section, we describe the data collection process, our data cleaning process, and some exploratory data analysis.

### 2.1 Data Collection and Description

Sensors, as part of a wireless sensor network, were placed on a 70 meter tall redwood tree approximately 2 meters apart from each other starting at 15 meters high (where the canopy began). Most sensors were placed on the west side of the tree as it had a thicker canopy and thus offered better protection against environmental factors not associated with the tree. Measurements from each sensor were taken 5 minutes apart from April 27, 2004 at 5:15pm PDT to June 10, 2004 at 2:00pm PDT. Important climate variables were collected, including temperature, humidity, and light levels. Temperature was measured in degrees Celsius and relative humidity was presented as a percentage. Light was measured using readings of photosynthetic active radiation (PAR) with a wavelength range from 350 to 700 nm (the range of light plants use to grow via photosynthesis) in units of micromoles per meter squared per second ($\mu$mol/m$^2$/s). Both incident (with sensors in direct sunlight) and reflected (with sensors being shaded) levels of PAR were obtained.

A brief description of each recorded variable and its associated units is provided in the table below.

| Variable | Description | Units |
|---|---|---|
| meas_time | Time the measurement was taken | YYYY-MM-DD HH:mm:ss PDT |
| result_time | Time the measurement was downloaded | YYYY-MM-DD HH:mm:ss PDT |
| epoch | Identifier for time the measurement was taken | |
| nodeid | Identifier for sensor | |
| parent | Parameter for sensor network structure | |
| depth | Parameter for sensor network structure | |
| voltage | Volts measured from tree | V (volts) |
| humidity | Relative humidity | %RH (relative humidity) |
| humid_temp | Temperature | °C |
| humid_adj | Adjusted relative humidity | %RH (relative humidity) |
| hamatop | Incident PAR (sensor on the top) | $\mu$mol/m$^2$/s |
| hamabot | Reflected PAR (sensor on the bottom) | $\mu$mol/m$^2$/s |

| Variable | Description | Units |
|---|---|---|
| Height | Vertical distance of sensor to the ground | m (meters) |
| Direc | Direction sensor faces on tree | North(N)/South(S)/East(E)/West(W) |
| Dist | Radial distance of sensor to trunk of the tree | m (meters) |
| Tree | Sensor tree location descriptor | interior / edge |

## 2.2 Data Cleaning

Before describing the details of our data cleaning process, we summarize it in the table below.

| Number of Samples | Data Processing Step |
|---|---|
| 416,036 | Initial Sample Size |
| 416,035 | Remove sample from node id (65535) for readings out of range |
| 416,034 | Remove sample from parent id (2058) for readings out of range |
| 395,406 | Remove duplicate samples |
| 383,326 | Remove samples with missing values across humidity, humid_temp, humid_adj, hamatop, hamabot |
| 255,264 | Remove samples with voltage outside the range 2.4 - 3.0 volts |
| 250,857 | Remove samples if the Reflective PAR is greater than the Incident PAR |
| 250,739 | Remove samples if they differ by $< 2$ units across any continuous variables |
| 250,729 | Remove samples if there exists a parent or depth mismatch |
| 250,723 | Remove samples if PAR readings are inconsistent |
| 250,721 | Remove samples if they differ by $> 2$ units across any continuous variables |

First, we verified that the tree data set in 'sonoma-data-all.csv' was a concatenation of 'sonoma-data-log.csv' and 'sonoma-data-net.csv'. Then, we attempted to match the 'ID' column from 'mote-location-data.txt' to 'nodeID' in the 'sonoma-data-all.csv' data set because, theoretically, each mote (collection of sensors/sensor unit) should correspond to a sensor node. There were 3 node IDs in the tree data set that did not match to a mote ID (including one node that was dropped later as part of data cleaning). As there was no other clear identifier to match with and with the majority of the node IDs matching, we stuck with this mapping and merged the two files.

Looking closer at the unique node IDs, we noticed one of the node IDs is '65535', which is equal to 2^16-1 (a common overflow number in computation). Taking a closer look at the readings for this node, it was clear they were not in sensible ranges for multiple variables and it was dropped.

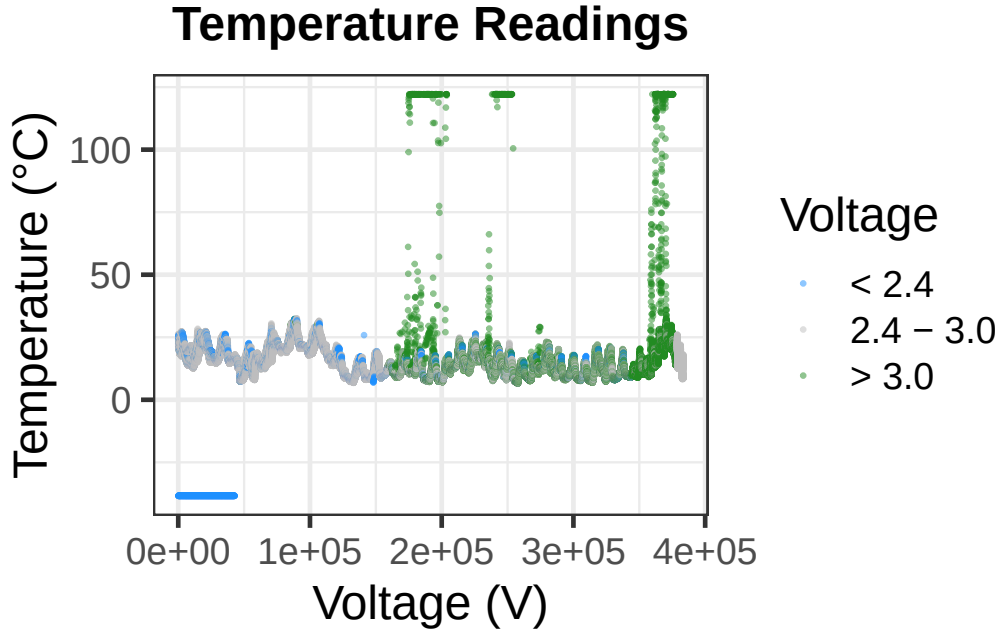
We next looked at the other network parameters, 'parent' and 'depth'. We checked whether the unique entries of 'nodeid' appeared in 'parent' (they did) but not all unique entries of 'parent' appeared in 'nodeid'. Two unique parent entries '65535' and '2058' also appeared to be out of the typical range of values which otherwise range from 0 to 200. We note that readings for nodes associated with parent '2058' had anomalous readings and were dropped. We verified that all 'nodeid' entries corresponded to a single 'parent'. Finally, we noted that 'depth' had 8-bit values and did not perform any cleaning beyond ensuring consistency between 'nodeid' and 'depth' across readings.

Next, we looked at the timestamps in the dataset. We saw that each entry of the 'result_time' column was most likely the time the data was downloaded and not when the sensor actually took a reading in real time. This appears to be the case because the time frame of the readings (May 7, 2004 at 6:24pm - November 10, 2004 at 2:25pm) are inconsistent with those of the study as described in Tolle et all (April 27th 2004 at 5:10pm - June 10th 2004 at 2:00pm) and the readings are also not in multiples of 5 minutes past the hour like they should be. Even more so, for all 301,056 samples in 'sonoma_data_net.csv', the time reading is uniformly the same date (November 10, 2004 at 2:25pm). However, looking at the data in 'sonoma-dates-epochDates.txt' and 'sonoma-dates-epochNums.txt', we found a list of specific measurement times and epochs (a number

associated to each time), respectively. We found that epochs and measurement times mapped to each other uniquely; hence, we decided to map the 'epoch' column in the 'sonoma-data-all.csv' file to the times in this epoch reference data instead of using 'result_time'. After doing this, the time frame of readings then very closely aligned to what was in the paper (April 27, 2004 at 5:15pm - June 10, 2004 at 2:00pm): all readings were in 5 minute intervals, and all epochs in the data had a match to this epoch reference.

After sanitizing the timestamps, we deleted any duplicate rows or rows that had missing values across all the main variables in the tree data ('humidity', 'humid_temp', 'humid_adj', 'hamatop', and 'hamabot').

We next looked at voltage values from the sensors. The authors of Tolle et al mention that poor battery life corresponds to strange voltage readings, and that any readings outside of 2.4 - 3 volts can not be trusted. We can see this clearly when looking at the different variables. Shown below is temperature; in particular, we can see that physically improbable values (such as those below zero or those that are very high near 100 C) are all coming from sensors with anomalous readings. Having this knowledge about the sensor's range and seeing how it leads to odd readings, we decided to stick with the range the authors used and remove volts outside the given range.



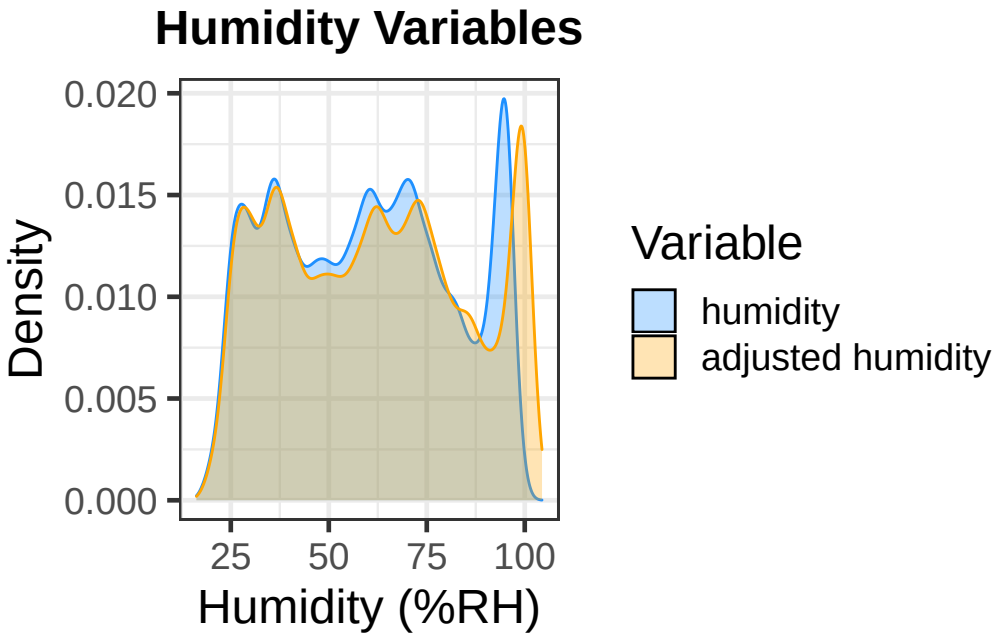

We also checked whether Reflective PAR readings were ever higher than Incident PAR readings, as this should never happen (reflected light must by definition be weaker than direct light). We removed samples for which this occurred as their measurement readings likely cannot be trusted.

Next, we noticed that there were sometimes multiple samples for the same (node, epoch) pair when there should only be one sample per (node, epoch) pair. We noticed this happened in several different ways. Occasionally, all readings for a node were the same except for voltage. Other times, there were inconsistencies between parent, depth, or both. And, lastly, there were sometimes inconsistencies between the PAR readings. Sometimes, there would be inconsistencies across the continuous variables but by at most 2 units (except for the node with identifier 3 at 2004-11-10 14:25:00, which has starkly different humidity readings: 85.39 and 48.14 percent). In the absence of further information or domain expertise, we have no way of knowing which of the samples is correct. Moreover, the misalignment of node metadata (parent and depth) is concerning. Hence, we chose to drop all implicated samples.

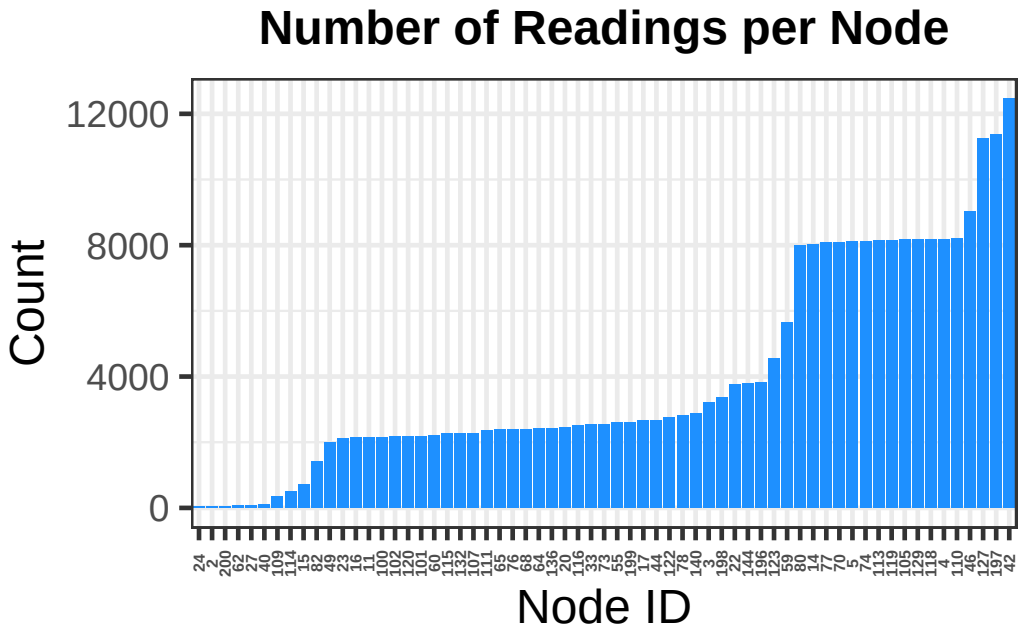
### 2.3 Data Exploration

There are two humidity variables in the data. As we used humidity in our analyses, it was important to understand what the relationship between the two is. From the figure below, we see that 'humid_adj' is just the humidity variable adjusted. Moreover, looking at the summary statistics of each, we see that the

maximum value of 'humidity' is 104.41 and that of 'humid_adj' is 100.2 (also the maximum value in Tolle et al). We suspect that 'humid_adj' is a scaling of the humidity readings to correct for various factors and to provide a maximum reading of 100. Hence, we use 'humid_adj' as our sole humidity variable in our analyses.



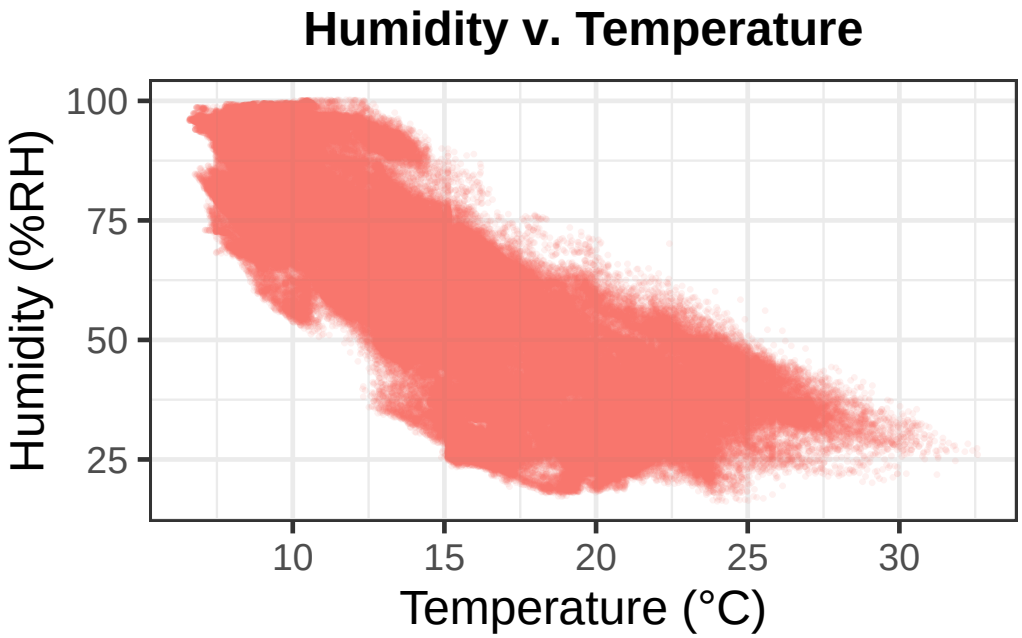

We then generated a plot of number of readings per node. Per the experimental design, they should all be the same but it is clear that they varied wildly across the nodes. We will come back to this and dive more into whether node lifespan is correlated with any other variables in the findings section.




## 2.4 Reality Check

From basic meteorological knowledge, we know that humidity should be inversely proportional to temperature; indeed, we see that it is. Moreover, we looked at summary statistics of the variables and saw they were in the correct ranges (except for 'humid_adj', which was a bit off as its maximum is 100.2, not 100). Finally, as we will see in a later section (Finding 2), the light readings exhibit a daily, cyclic pattern. We note that the

majority of our sanity checking occurred as we cleaned the data, e.g., the node-epoch pair uniqueness and epoch-timestamp correspondence verification.




# 3 Graphical Critique

## 3.1 Figure 3

The goal of Figure 3 is to understand how temperature, humidity, and both light readings vary on their own as well as functions of time and height.

In the subfigures in 3(a), we see histograms of the readings. As an overall view of what is happening, this is helpful for gaining a sense of what typical values for measurements are. One point to note here is that all figures appear to use the same number of bins: it is likely that a technique based on sample counts was used to pick the number of bins, or, more simply, a number of bins that led to readable plots was used.

In the subfigures in 3(b), we see box plots of the values over time. That is, for each day in the study, there is a box plot. This figure is quite crowded, and it is hard to get a sense for what is happening. Granted, we must note that the boxes form something of a band so that we see a band of values across time; however, the incident PAR plots in particular have long tails, and it is hard to get much out of the plots. A trend line, or, a truncation of the tails in the box plots might have been helpful. A similar complaint and summary can be made for the subfigures in 3(c), where for each height value a box plot of the readings is given. If all of these were larger, some of these issues would be ameliorated. Note that the idea of these plots is good, and the data could be presented well using this method with some further refinement.

In the subfigures in 3(d), box plots for differences from the mean (over time) are plotted for each reading as a function of height. That is, in the same style as 3(c), we have horizontal box plots but this time of the differences of sensor readings from the mean. The same issues as in 3(b) and 3(c) persist, namely that many of these box plots have long tails and the plots are crowded and hard to read. The goal of these plots was to find trends as a function of height; it is hard to see where zero is on the x-axis, but even so, the PAR/light plots are reasonably readable.

## 3.2 Figure 4

The goal of Figure 4 is to understand temporal trends within a single day.

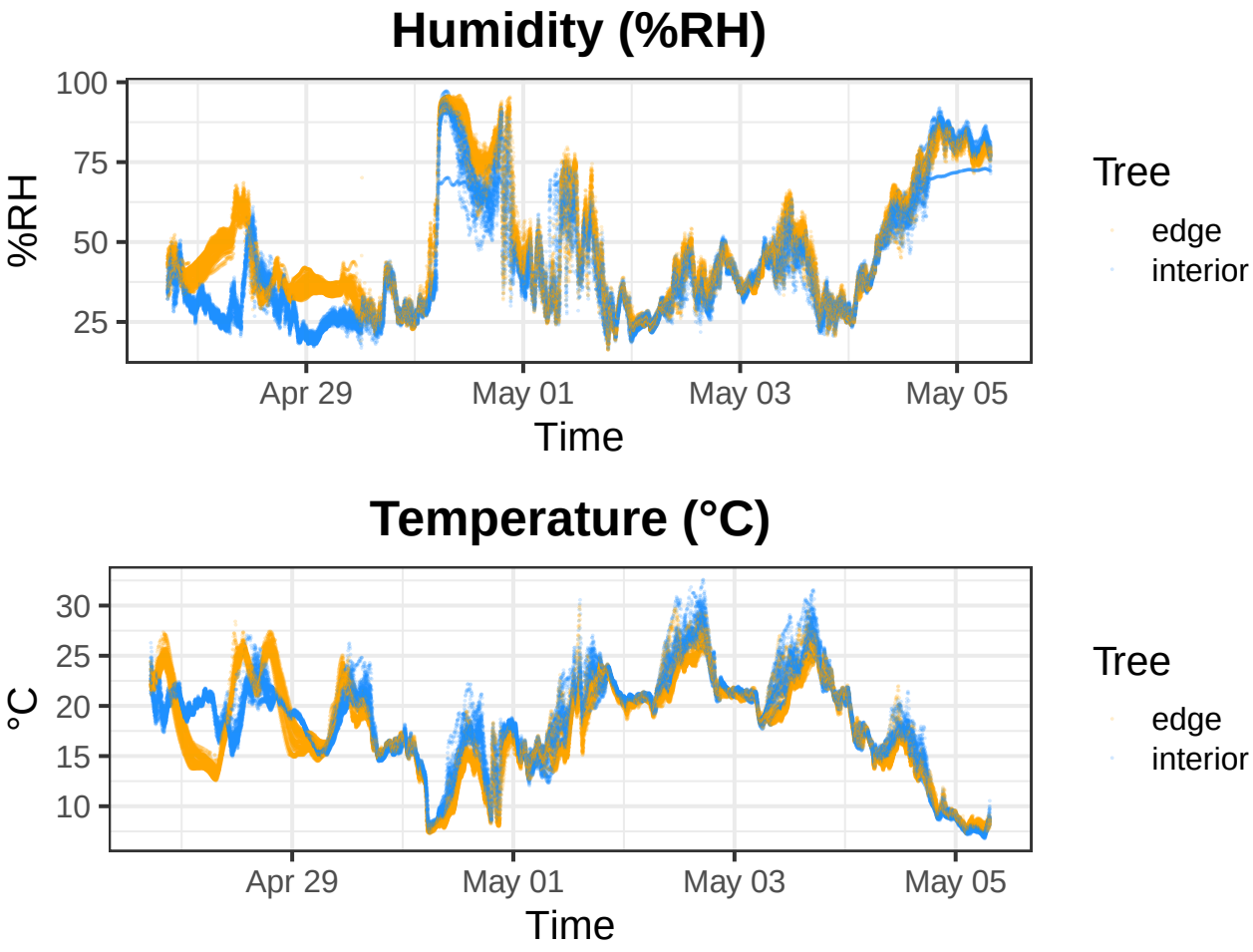
In the first column, the first two rows plot temperature and humidity over time for each node. We see that all of the lines vary together within each plot. It might have been both cleaner and better looking to show a mean trend line and bands around it to capture the same effect, rather than dozens of lines all superimposed. The third and fourth rows in the first column repeat the same idea but for incident and reflected PAR (light). Rather than lines, there are dots plotted and a solid trend line/average. It would be good to know how the trend line was computed: looking at the density of points, it seems as if it is pulled upward of where the bulk of points are by relatively few high-valued points. It is better that there are points here and not lines, as the plot looks cleaner. Note that the choice of green to indicate sunlight is somewhat unusual, and something redder in color might have been more thematic. That said, all of these plots are effective in conveying their message and a sense of what the daily variation in the tree is like.

The second column of Figure 4 is somewhat difficult to understand, and I found the description in the text to be somewhat difficult to follow. For a particular time in the day, the goal was to present the variation in the readings as a function of height. Each plot has a trend line and some plots around it. The trend line is in a faint green color, and is somewhat hard to follow. There are triangles with different colors (blue and pink) that indicate gradients, where the gradients are taken at the time indicated at the vertical line in the first column of Figure 4. The final plot (for reflected PAR) is a poor use of space: the majority of the x-axis is devoted to empty space and it is hard to see anything. Granted, it is possible that this plot was intended to present the lack of variation, as it has the same range as the corresponding plot in the first column.

# 4 Findings

## 4.1 First finding

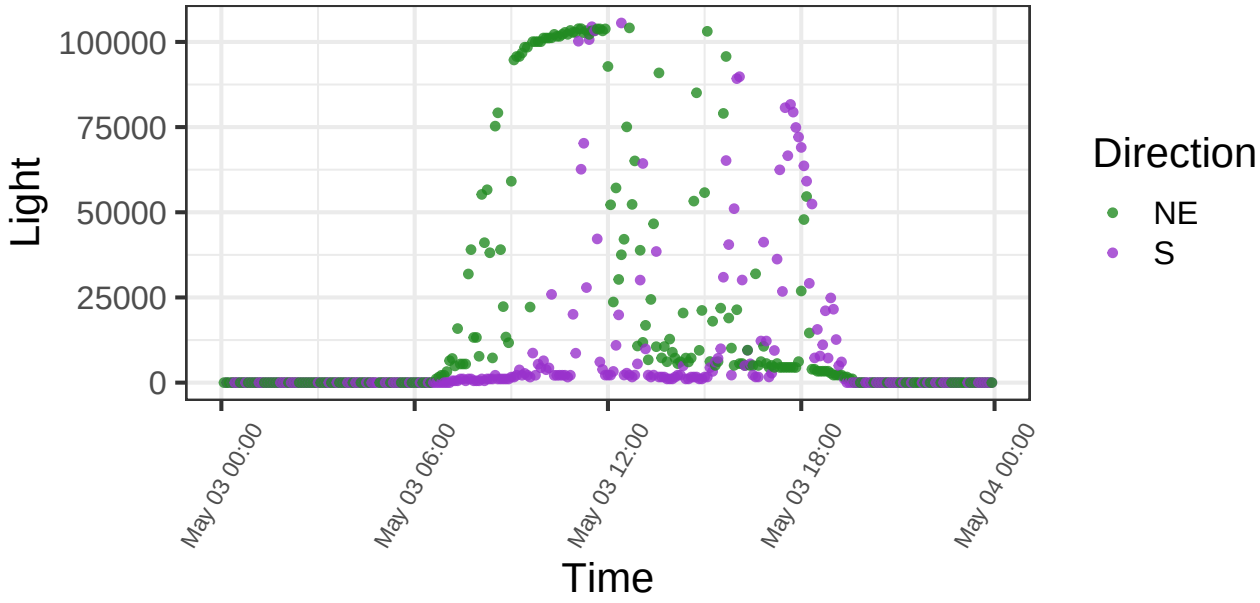
We investigate the effect of the tree on temperature and humidity, that is, whether the tree is an interior or an exterior tree. We plot the temperature and humidity for the period April 27 to May 05, grouped by interior and exterior tree status. Note that we only have both of these labels for this time range, as there are no 'exterior' readings after May 05. In general, we see that the readings are similar, i.e., that there is minimal variation. We conclude that to capture the dynamics of a redwood forest, it is most likely not necessary to sample adjacent trees, but a wider spacing/lower sensor density would most likely suffice.



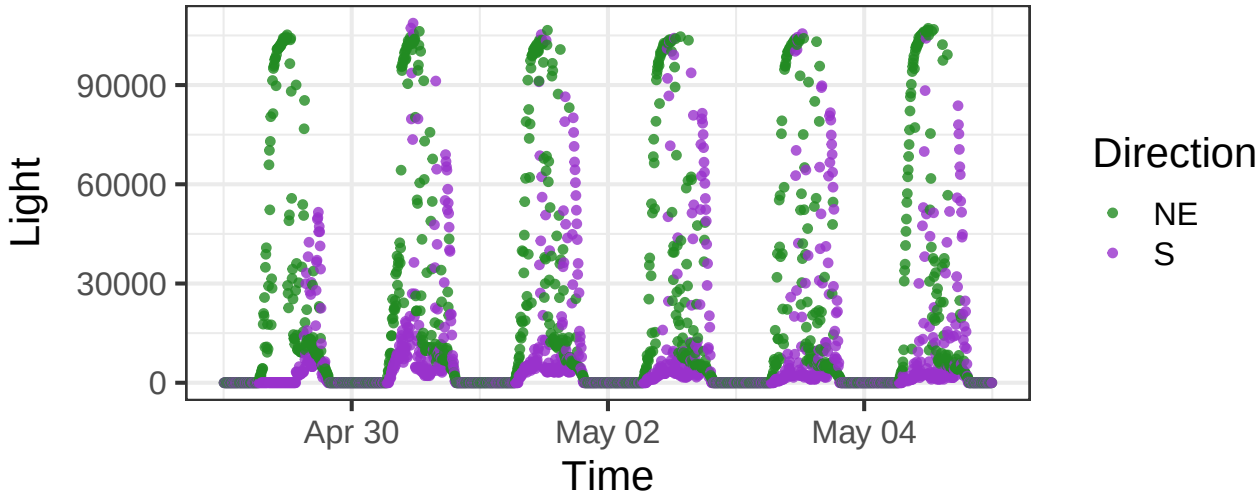

### 4.2 Second Finding

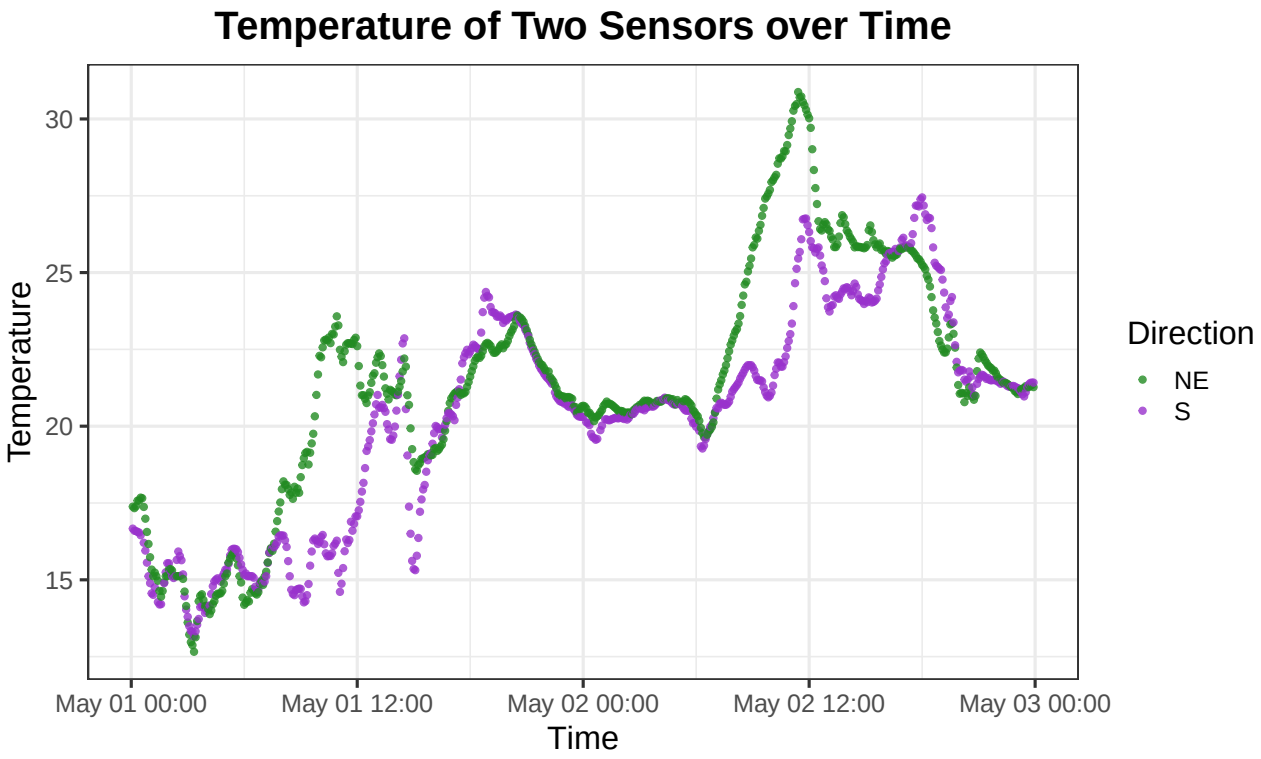
We study the relationship between light (Incident PAR) and direction across time. We focus on two diametrically opposed nodes, one in the northeast of the tree and the other in the south. In particular, we look for a lagged response between the two nodes: as the sun passes through the sky, the northeast node should have an earlier peak the the southern node. Moreover, knowing that in the northern hemisphere, the south and west directions tend to receive more/stronger sunlight, we looked to see if there was a significant difference between the peak levels. We present two figures, first of the light readings on May 03, and second of the light readings between April 27 and May 05. We see that the northeast node peaks a few hours before the southern node. Moreover, our hypothesis about light levels is not substantiated, as the readings between the nodes were generally similar, except for on May 02 and early on May 03, where the northeast node had a marginally higher peak. We also look at temperature on May 01 and May 02, and see that the temperatures on the northeast node peak before those on the southern node. It is interesting to see this similar lag structure between the directions, and also to see that the northeastern temperature peaks at a higher value than that of the southern node.

## Incident PAR of Two Sensors on May 3rd



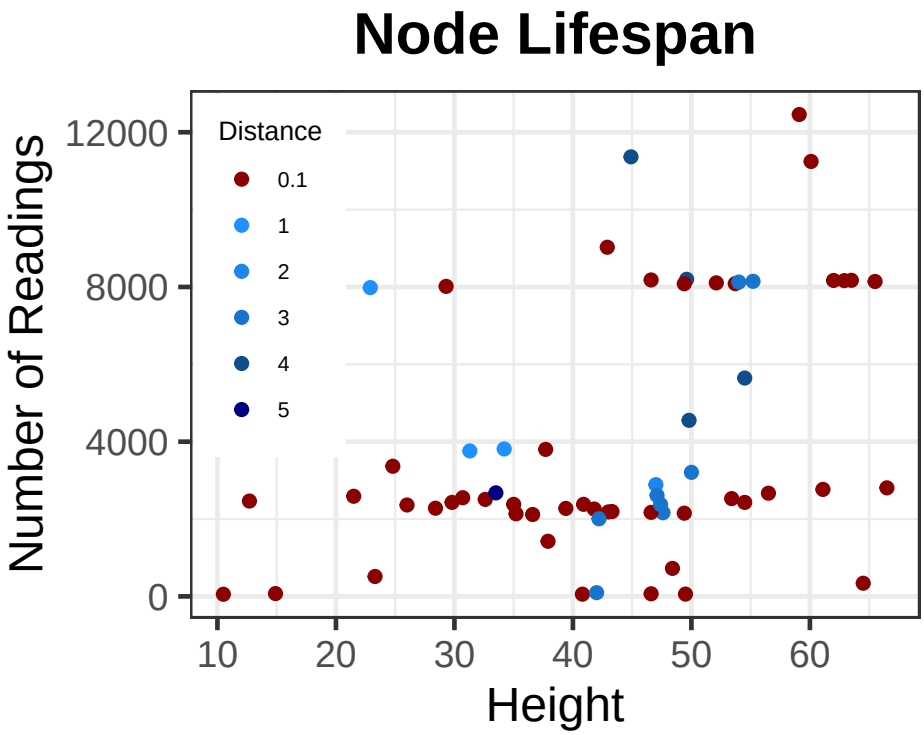

## Incident PAR of Two Sensors over Time

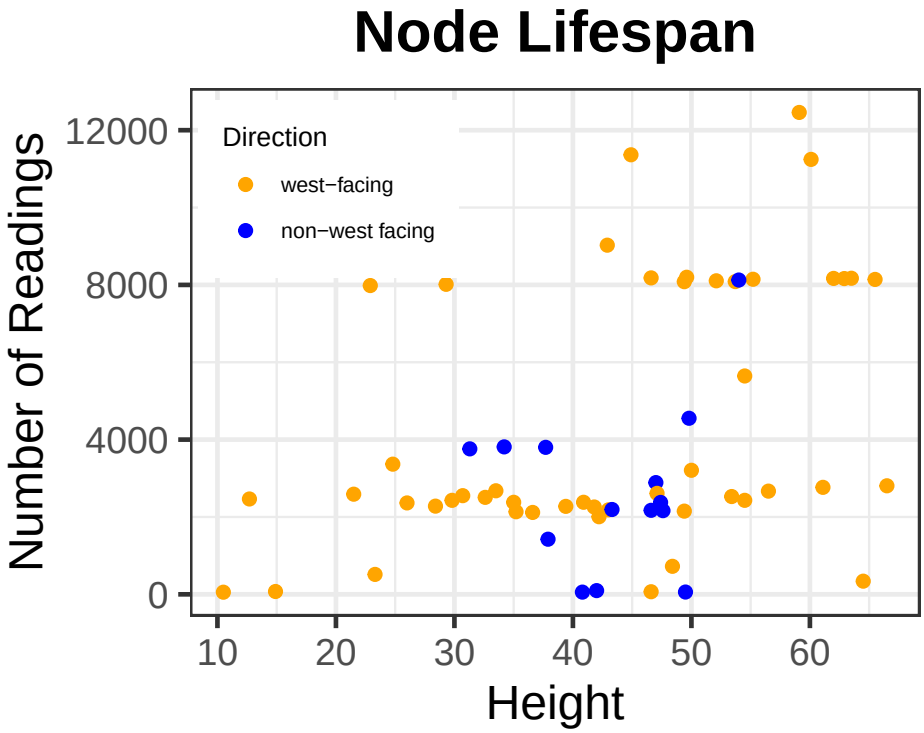





### 4.3 Third Finding

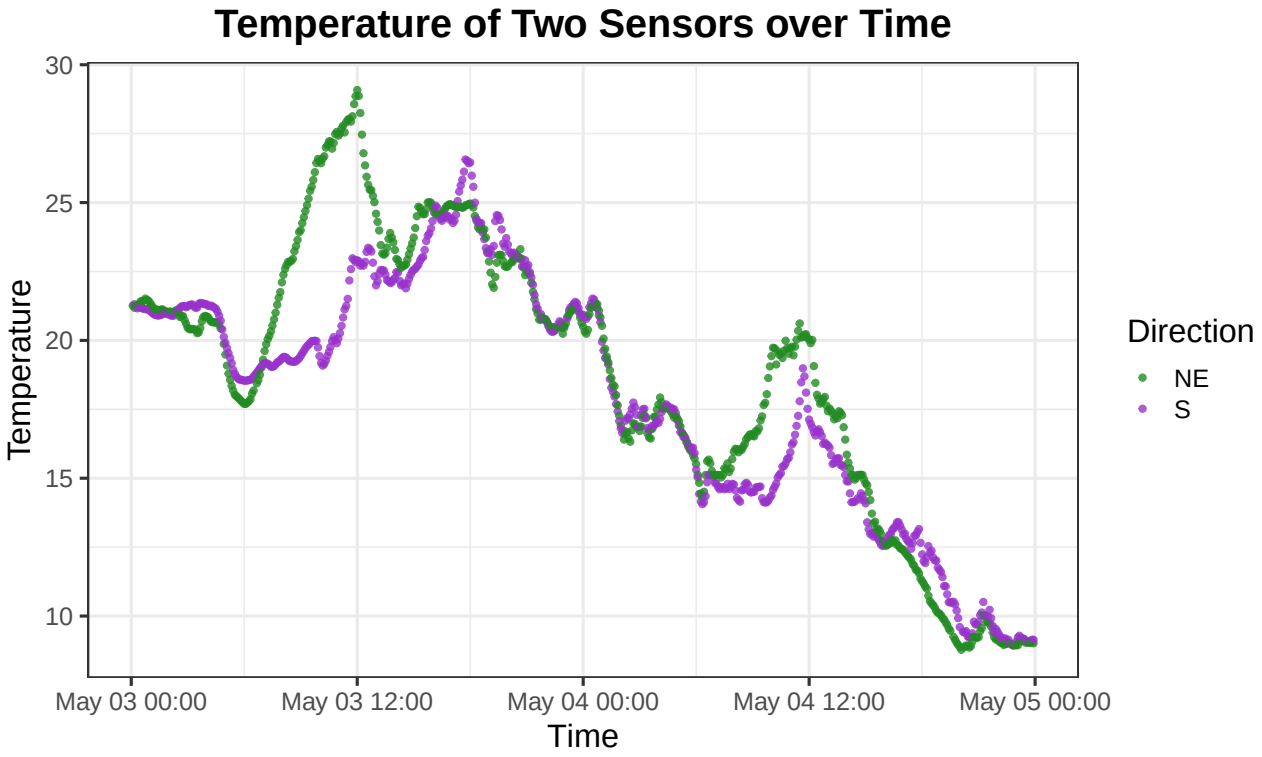
Recall that in Tolle et al, they placed the majority of the nodes in the west and south-west directions, primarily to protect against environmental factors. Furthermore, nodes were placed at varying distances from the trunk, but were primarily placed close to the trunk, so that measurements characterized the tree and not its vicinity. In our data cleaning and exploration process, we noticed that several nodes had missing readings or readings that had to be dropped. It is plausible that environmental factors, perhaps related to directionality, distance from the trunk, or height (and hence shelter from the elements), impacted the longevity or the number of readings that a particular node yielded. We present two figures below. In both, we look at the number of readings as function of node height. In the first figure, we color the points by distance from the trunk, and in the second, we color the nodes by their direction. We see that there is no clear relationship between the factors, or, that these plots do not indicate that node placement affected node lifespan.




### 4.4 Stability Check

We return to finding 2, where we select a different time window and return to the temperature analysis. We keep the directions identical (northeast and south). Indeed, we see a similar result as in finding 2, where the northeast temperature peaks earlier than the southern temperature by a few hours, and at a higher value once again. Note that the pattern, while continued, is not as strong as before. Ideally, we would study this for more node-direction pairs and across a larger time window to substantiate this relationship.




## 5 Discussion

We have performed an extensive data cleaning procedure on this dataset. Indeed, after cleaning, we are left with 60.2 percent of the original data. While there are thousands of samples, there are relatively few covariates, so that the data size is not problematic to handle. However, the large number of samples makes visualization tricky, as we must restrict our view to smaller windows or a subset of nodes to create meaningful and interpretable plots.

Along the way, we made several assumptions about the data and its structure. Many of these assumptions were borne out of necessity: it is difficult to analyze a dataset that we have not designed and cannot query the collectors about. In particular, many of the samples that were discarded in cleaning were discarded because of metadata issues, e.g., node id, timestamp, and metadata fields like 'parent' and 'depth' not matching up. Moreover, we made assumptions about the relationship between the humidity and the 'humid_adj' fields in the data that we cannot fully substantiate. It may be necessary to perform further exploration of this variable before further analyses. Another variable whose processing relied on assumptions is voltage. While we followed the prescription in Tolle et al, it is not clear whether this is the best path forward.

This dataset was collected in a real forest on real trees. That is, it offers a view of reality, albeit one that is necessarily skewed by how it was collected, the various events that occurred during its collection (sensor failure, data storage overflow, etc.), and how it was stored and processed. Moreover, our visualizations necessarily carry our biases and further color the view of reality in the data. However, it still offers great potential for effecting a change going forward. This line of research can help understand how redwood trees and forests are affected by climate change, as well as how their microclimates behave.

A longer term study with more trees and more robust sensors would be a natural next step. Such a study would allow for a view into the behavior of the trees across seasons, and would offer more, hopefully cleaner and higher quality data to analyze.

# 6 Conclusions

We have studied the dataset collected and presented in Tolle et al. We have performed an extensive data cleaning procedure that found several inconsistencies in the dataset, a brief exploratory data analysis, and a short look at a few interesting questions about the relationships between the variables. Our analyses look at questions about the study itself (the effect of node/sensor placement on data) and about the underlying natural processes (e.g., temperature v. humidity, sunlight as a function of direction) that directly address the main problem of understanding the microclimate of a redwood tree.

# 7 Academic honesty statement

Professor Bin Yu,

All coding for the analyses, written text for the report, and other associated work for this lab was done by me and represents my own work.

# 8 Bibliography

A macroscope in the redwoods, *Tolle, Gilman and Polastre, Joseph and Szewczyk, Robert and Culler, David and Turner, Neil and Tu, Kevin and Burgess, Stephen and Dawson, Todd and Buonadonna, Phil and Gay, David and others*, Proceedings of the 3rd international conference on Embedded networked sensor systems.