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

*blank*
*25948127*

*September 9, 2017*

## 1 Introduction

This report details replication of the analyses in Tolle et. al.'s paper, "A Macroscope in the Redwoods" (2005). In the initial experiment, the investigators placed 72 remote sensor nodes on a California redwood tree, at various vertical positions. Each sensor recorded information about the surrounding air temperature, relative humidity, and photosynthetically active solar radiation. In addition, the raw dataset included covariate information about the sensors themselves about their position, number of records, and battery voltage.

The general goal of the initial study was to examine and describe the manifestation of microclimates in the area immediately surrounding the redwood tree. The tree is 70 meters (m) tall, and the investigators suspect that there is notable variation in things like humidity, temperature, and sunlight across the height of a redwood tree. Historically, this variation was difficult to measure directly because simple instruments cannot collect and store data in regular intervals over a long period of time, and human beings are not equipped to observe this sort of variation consistently. However, the authors deploy remote sensors that, over the course of 44 days, measure these features.

In terms of this report, most of my findings and discussion focus on the quality of the **measurement instruments**. Although the paper itself largely elides this topic, I found the remote sensor nodes to be the source of the most interesting anomoloies in the data. I also perform some general exploratory analysis on the relationships between different variables. My general findings are:

- Battery death was not random, and was possibly related to humidity, which skewed the records
- "Edge" and "Interior" nodes form two distinct networks with relatively weak connections between them
- The top of the tree got far more sunlight than the bottom

This report proceeds as follows: Section 2 describes the data source and collection, Section 3 describes a critique of the graphics presented in the original paper, Section 4 presents my own original findings from the data, Section 5 offers a discussion, and Section 6 concludes.

## 2 The Data

### 2.1 Data Collection

The data were collected by deploying 72 sensor nodes along the trunk and branches of a redwood tree. They were placed on the west side of the tree as this side was more sheltered from weather. The data were collected about every 5 minutes across a 4 month span.

The raw data are confusing at first because explicit dates are only cover about a three week-long span between May 7th, 2004 and June 2nd, 2004. The rest are given a date of November 10th, 2004 and time of 14:25:00. There are also other odd measurements such as negative humidities, and implausiblly large photosynthetically active radiation (PAR) measurements, which suggests the sensors were prone to problems. Thanks to supplementary date information, I was able to correct the date and time information and construct a full dataset.

In my view, the most interesting variables were the covariates describing the remote sensors. In particular, I focused on voltage, vertical placement on the tree ("Height"), and whether the node was placed on the

interior or exterior of the tree ("Tree"). Otherwise, I also used the humidity, temperature, and epoch (time periods) to explore the data and draw conclusions about the validity of the nodes' inferences.

## 2.2 Data Cleaning

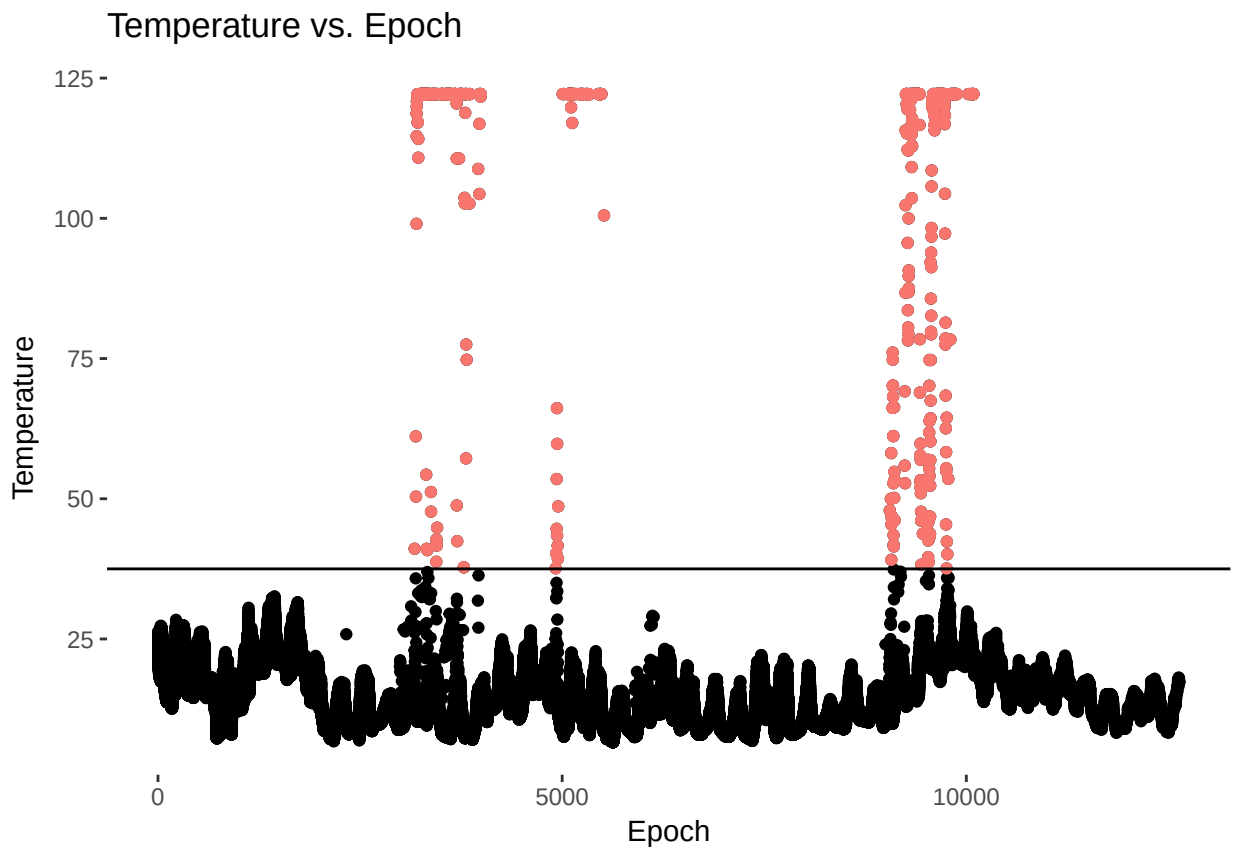
There are several data cleaning steps necessary before proceeding with analysis. First, I used the supplied "load" and "clean" functions to load in correct Date-Time information, the measurements datasets, and the node information. I edited the "load.R" file to use tidyr/dplyr functions because of issues with Windows compatibility. Next, I joined the redwood dataset, correct dates/epoch numbers, and mote data together into one master dataset (which I titled "redwood_master" in the source code).

I also suspect that some of the voltage data records were incorrect. While some were clearly recorded in volts, others recorded values in the hundreds (which is implausible for the AA batteries used). I suspect that these were recorded as centivolts for some reason. I cleaned the data by dividing any voltage values that were in the hundreds by 100, and recorded the new adjusted voltage column as "voltage_adj." However, these values look distributionally different because many of them fall below 2.5V, so I am not 100% confident in my guess.

At any rate, each row now has information about DateTime, Date, Time, epoch number, node ID number, parent node, voltage, depth, humidity, temperature, adjusted humidity, treetop sunlight, reflected sunlight, vertical placement on tree, direction of the node, distance from trunk, whether the node was on the interior or exterior, the day of week, and adjusted voltage.

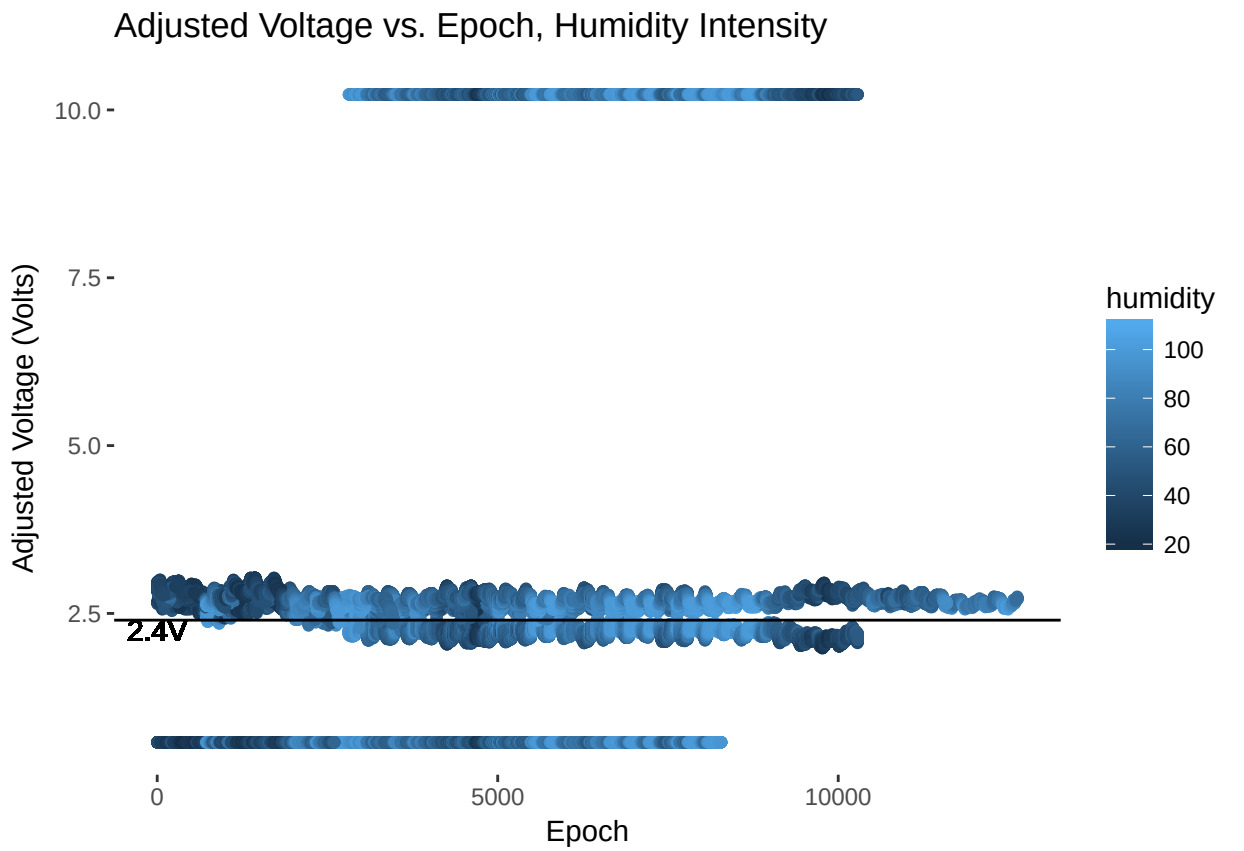
Otherwise, I also dealt with outliers, but am concerned about the method that was used in the paper. First, a simple graph of the temperature on humidity reveals that there are a few points with humidity readings less than 0. Obviously this is impossible, so I discard those. I do not plot this because of space constraints, but the plot is available in my additional R code.

I also plot temperature on epoch to see if there are implausible temperature values. Here we see several outliers that are well above the 30 degree Celsius upper range that we would expect in a redwood forest. In fact, there are a non-trivial number of observations in the boiling range. Here I color (in red) in observations equal to or above the 37.5 C mark, which is still quite generous as this sort of observation is in line with the hottest days *ever* recorded in the redwood forests. Because these measurements above this line were implausible, I exclude them as well.




Next, I turn to the original paper's data cleaning method. The authors correctly note that measurement errors are correlated with battery failure which is defined as battery voltage dropping below 2.4V. In my reading of the authors' explanation, they seem to assume that these failures happen basically at random. However, I suspect that this is actually *not* the case.

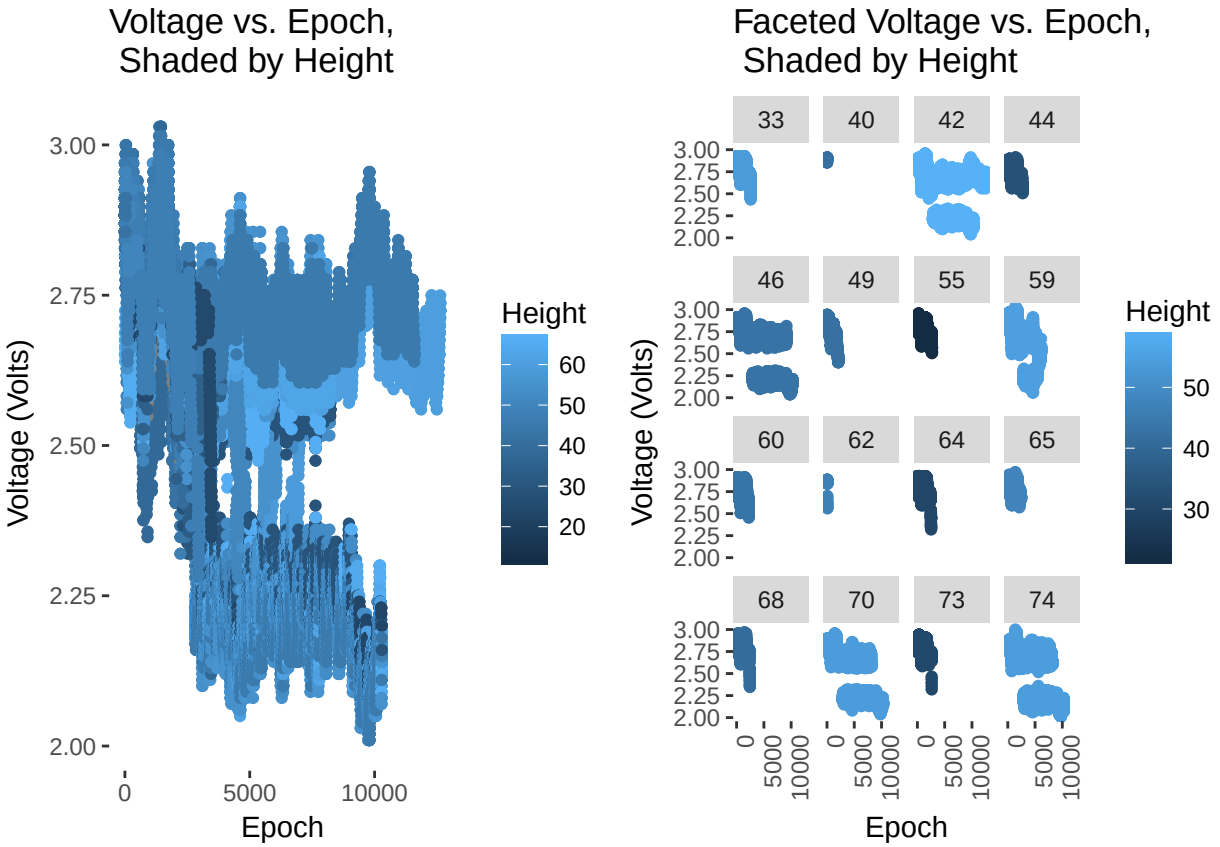
I detail my exact work in the findings section below, but basically, I suspect that humidity might increase the probability of battery failure. As is plain to see, assuming my suspicion is correct that some of the observations were recorded as centivolts instead of volts, a good portion of the data falls below the 2.4V mark, but still above the 2.0V mark. After reviewing the device information, I hypothesize that in these devices, one of the AA batteries depleted rapidly, but the device was still somewhat operational. Each AA battery contributes 1.5V to the total 3.0V at the start of the device's lifetime, so one suddenly rapidly losing charge could bring down the overall voltage below 2.4V without necessarily killing the device. So, unlike the authors, I choose not to exclude these data right away, but I will exclude data with voltage readings around 0 and above 10 as these indicate that the device probably never worked. Note that this graph was constructed after excluding extreme humidity and temperature readings.




## 2.3 Data Exploration

### 2.3.1 Exploring Voltage

These plots visualize the relationship between voltage and elapsed time, colored in by the height of the node. I explored several different relationships, but this turned up to be the most interesting because it suggested a non-random relationship between height and voltage. The forest floor tends to be more humid than the tops of the trees, and the graphs suggest that the batteries closer to the floor (and therefore exposed to humid conditions for a longer period of time) dropped below 2.4V more quickly (as indicated by the fact that only high-up nodes survived to the end of the study). The faceted charts (of a sample of the nodes) also demonstrates this trend, as the lightly colored nodes (close to the top of the tree) tend to survive to the end of the study, and the dark-shaded ones die earlier on. Unfortunately, the plots are a bit visually unappealing. Because there are so many points that behave in a heterogenuous fashion, the visual suffers from overplotting. However, the general trend should still be visible.



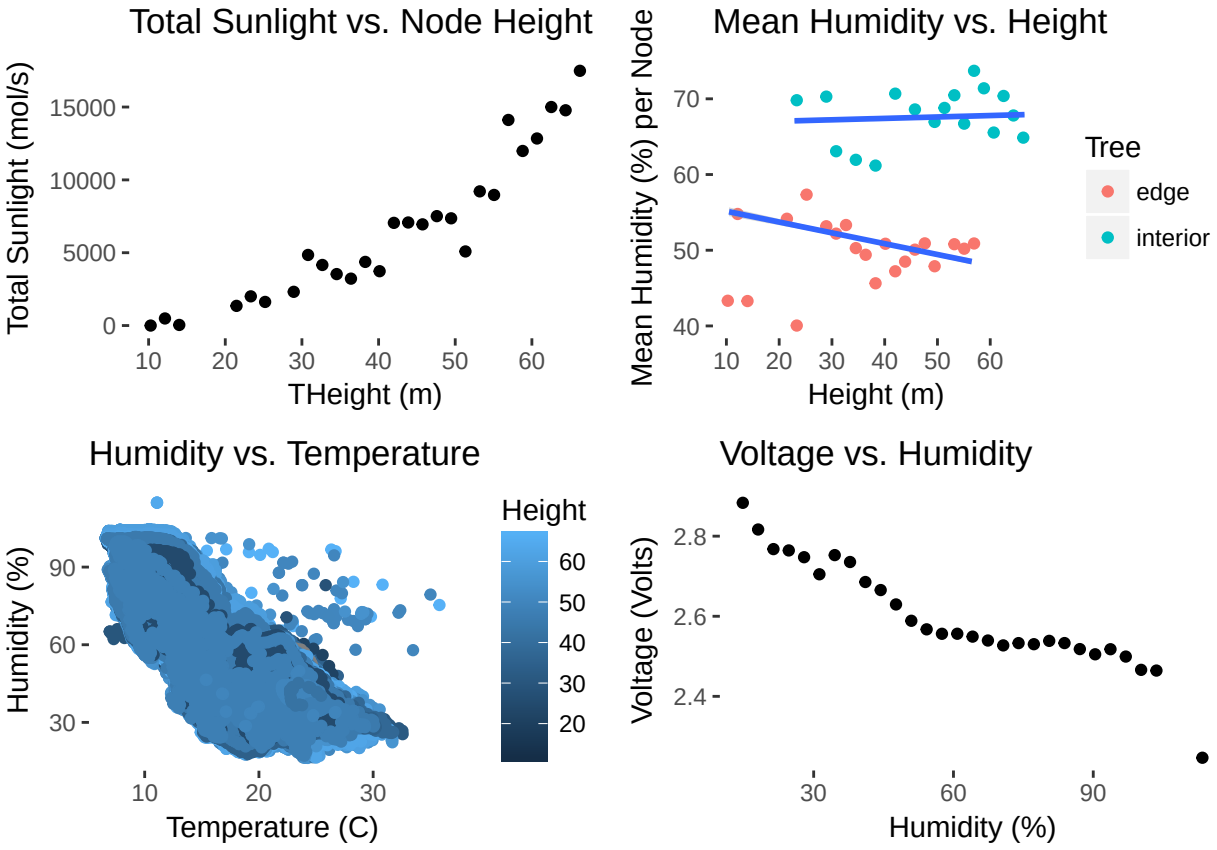

### 2.3.2 Exploring Height

I also explored the relationship between humidity and the other variables. The authors' were primarily concerned with the interplay between sunlight, humidity, and temperature, so exploring the visualizations describing these relationships helps illustrate their motivation. On the top left, I looked at how sunlight was explained by a node's height on the tree. As expected, sunlight positively correlates with node height.

Next to the first plot, I plot the relationship between mean humidity and a node's height on the tree (and therefore the amount of sunlight it receives). When I initially plotted this, it looked like humidity rises with height. However, after breaking down whether the node was on the interior (blue) or edge (red) of the tree, this relationship effectively disappear. All of the "edge" nodes are below the 60 meter mark, whereas several of the "interior" ones are above that mark. Although there is no information about surrounding vegetation, the tree is only 70m tall, suggesting that some of thes "interior" nodes represent the very top of the tree.

Among the "edge" nodes, there is a negative relationship between humidity and height. The interior nodes may have also been better shielded from the elements regardless of position, which may explain why they survived the higher humidity at the same height as edge nodes. I do not reject some of the outliers because I do not have a good explanation for doing so, but even with them, there is evidence for a negative sloping relationship between humidity and height (meaning it is MORE humid the further closer to the ground the node is placed).

Below these, on the left I graph the relationship between voltage and humidity, given tree height. I subset only to those nodes in the "edge" category so as not to bias the results with the disproportionate number of nodes that were placed in the sunnier, drier part of the tree. There are overplotting problems here, but the negative slope and lighter gradient as temperature increases indicates that humidity decreases as height and temperature increase. Finally, on the bottom right, I plot average voltage versus humidity.




# 3 Graphical Critique

In this section, I turn to critiquing the graphs presented by the authors of the original study.

## 3.1 Figure 3

Figure 3 was confusing mainly because the authors stacked twelve different plots on top of each other that each contained a lot of information. In the first row, the authors plot relative frequency charts for temperature, humidity, incident PAR, and reflected PAR. I found this information to be uninformative because it was totally decontextualized from other interesting information. For instance, the relative frequency of different temperatures broken down by height would have been interesting, but by itself, it tells me very little. Indeed, the premise of the study is to analyze the microclimatic differences along the tree. The way these charts are plotted, they give an overly broad look at the data, and tell me nothing about interesting variation across strata. Whatever variation that they do display is likely masking underlying phenomena, which erodes their usefulness for understanding the experiment.

The second row has a different problem in that each graphic has a little too much going on without easily interpretable trend lines. The authors were clearly trying to show the change in the center of the distributions of the variables over the course of the study. Without drawing a trend line however, the reader is forced to eyeball the trend while looking at very small box plots. Because there are so many boxplots in a very small frame, they are difficult to interpret.

The third are fourth rows are a bit better in these regards as stacking the boxplots on top of each other gives the reader a sense of the spread of variables at different heights. However, I had a hard time deciphering

what each row was trying to convey. Visually, they look similar, but the third row is "projected onto time x value," and the fourth row is "projected onto height x value." This wording does not have a straightforward interpretation without reading sections 5.3 and 5.4, and even then, it is unclear what the authors mean precisely. While they say the information gives "temporal" and "spatial" information respectively, the axes are exactly the same and there is no reasonable indication as to how time or height enters the graph. In fact, both y-axes use "node height," making the "time x value" projection confusing. In my own exploratory analyses, I attempted to analyze node height, a variable, and time by plotting on x and y, and then shading the data points based on the third dimension. While this had its own problems with overplotting, it had a more straightforward interpretation than the authors' graphs.

### 3.2 Figure 4

The main issue with the plots in Figure 4 was the lack of a legend that could help distinguish between the various trends presented. For instance, in the "Temperature vs. Time" plot, I assume that each color represents a different node, but this is not explicitly stated anywhere. Moreover, without a legend, there is no way to tell which node is which, making it impossible to see any differences between low or high nodes. The vertical blue line similarly does not have a label, which makes it impossible to determine what it is supposed to represent (in actuality it is the start of the "day").

The smaller plots on the right similarly suffer from labeling issues. Without context from the text, there is no way to determine what points mean. The text explains that the graphs represent one day's worth of data, and places the data points at their proper height. However, placing height on the y-axis is a bit confusing as it is not the dependent variable. While the authors explained the reason, the choice does obscure usual statistical reasoning.

The last plot that shows the incident PAR over time also has the same problem, with the added issue of not labeling its trend line. Again, the authors explain this in the text, but the graph should make it explicit that the line is measuring the center of the distribution at a given time. Overall, the lack of titles, labels for multiple different data sources, and confusing axes make it tempting to gloss over these graphs and ignore the authors' general findings.

## 4 Findings

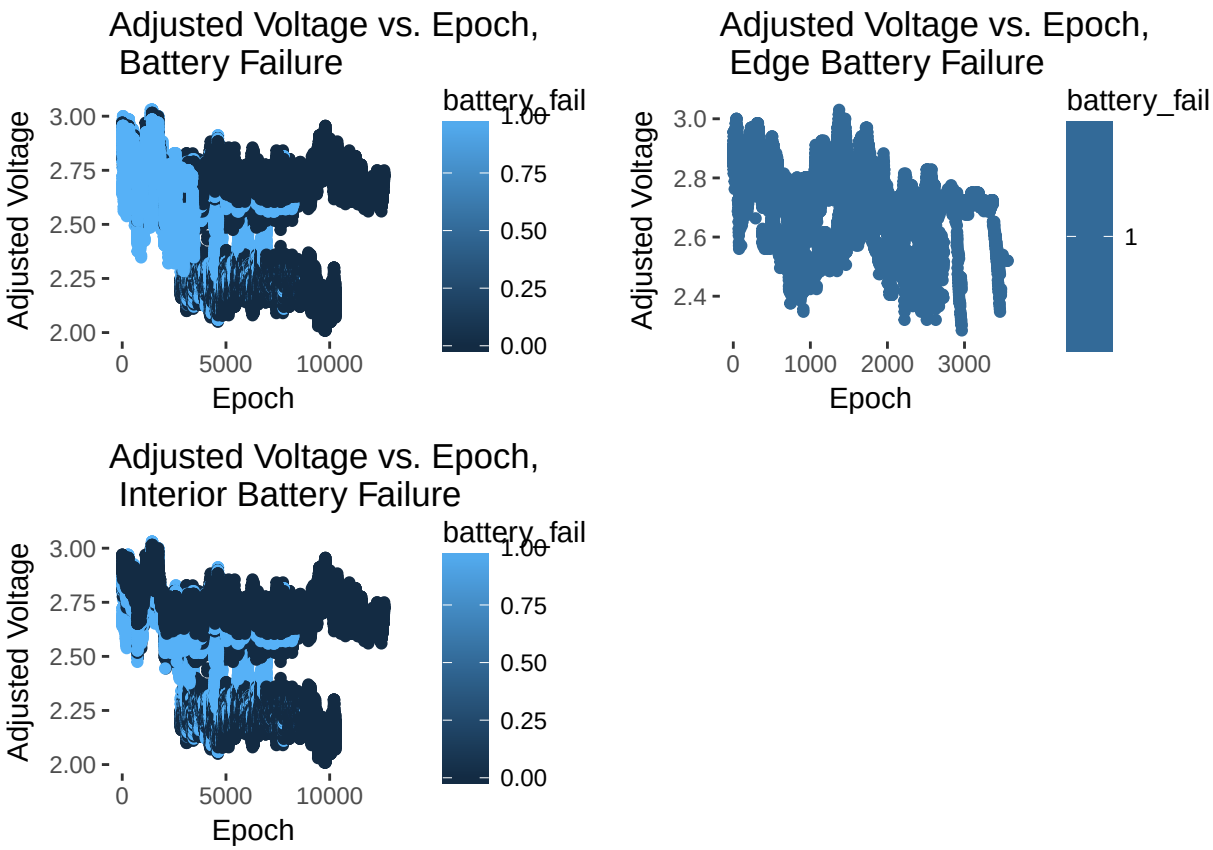
### 4.1 First finding

My first finding is that I suspect that remote sensor failure was not random. Instead, I hypothesize that the causal chain went something like: "sensor height on tree -> average temperature and humidity experienced by the sensor -> propensity for voltage drop -> battery failure." Specifically, I think that there is evidence to suggest that sensors placed lower to the ground experience more humid microclimates, and this humidity tampers with the battery's performance, which then can cause a battery to fail (and hence skew the measurements or stop taking measurements altogether).

#### 4.1.1 Battery Failure

Leveraging my exploratory analysis above, I suspect that there is a deep relationship between sunlight, temperature, humidity, and voltage. To test this for sure, I created a variable that measures battery failure. These graphs plot adjusted voltage relative to epoch, shaded by whether the battery failed before epoch = 10,000. I used this epoch as a rough proxy for a battery failing to make it to the end of the study. Although the study goes to epoch = 13,000, the number of sensor deaths drastically increases after 10,000, which makes it difficult to distinguish between relatively high performing sensors and sensors that died early on.

The most striking result here is that the graphs suggest that *no* edge nodes survived to the end of the study. In fact, all of the edge nodes died before epoch=3000, whereas a sizable fraction of interior nodes survived until the end of the study. The first graph shows the entire dataset, the second graph shows only edge nodes, and the third graphs shows only interior nodes. Light blue reflects battery death, and dark blue reflects a battery surviving. Overplotting is an issue here, but this was the best method I found for conveying this finding.




A probit model also lends support to the idea that a node's vertical location, exposure to humidity, and battery failure are correlated. Given that battery failure is not random, this suggests that the instrument was highly sensitive to moisture, and the authors' conclusions about the anything below the 60m is likely invalid. The estimate for regressing battery failure on adjusted voltage, humidity, Height, and Tree (interior or exterior) turns up with all independent variables being statistically significant.

```
##           term      estimate    std.error   statistic      p.value
## 1  (Intercept)  1.1097322338 2.810560e-03  394.843816 0.000000e+00
## 2  voltage_adj -0.0085171597 4.088279e-04  -20.833117 2.442955e-96
## 3     humidity -0.0014896818 2.235458e-05  -66.638780 0.000000e+00
## 4       Height -0.0002970425 5.031901e-05   -5.903186 3.568353e-09
## 5 Treeinterior -0.7950545782 1.559957e-03 -509.664346 0.000000e+00
```

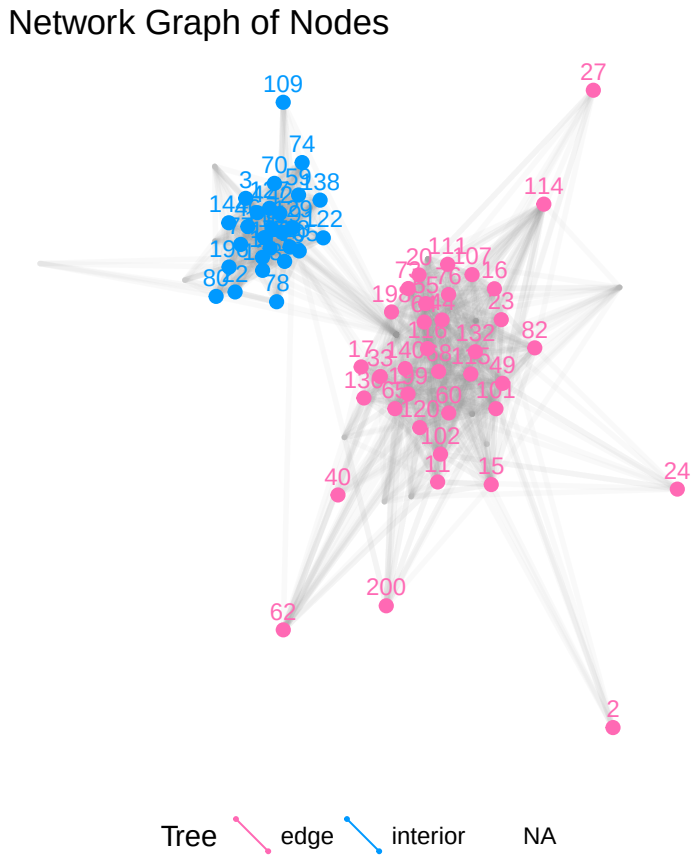
## 4.2 Second finding

My second finding further explores the relationship between edge and interior nodes. In particular, I look at whether edge and interior nodes detect each other, or if there was a systematic difference in which nodes were connected to other nodes.

To illustrate this, I construct a social network graph that shows the strength of connections between the

various nodes. Each point represents a particular node, and the thickness of a line between two nodes represents how many observations recorded a connection between the two nodes (where one was the child and the other was the parent). The closer a point is to the center of a group, the most numerous and diverse its connections are. The most interesting finding is that the edge and interior nodes effectively function as two separate networks, with a fairly weak set of connections tying the two together.

This result explains explain why I saw a noticeable difference in the voltage readings of edge and interior nodes above. The close clustering of nodes based on type, with few linkages to the other type, suggest that a general failure will be felt throughout the entire network. Because we have two networks, the edge nodes being so interconnected explains their collective failure.

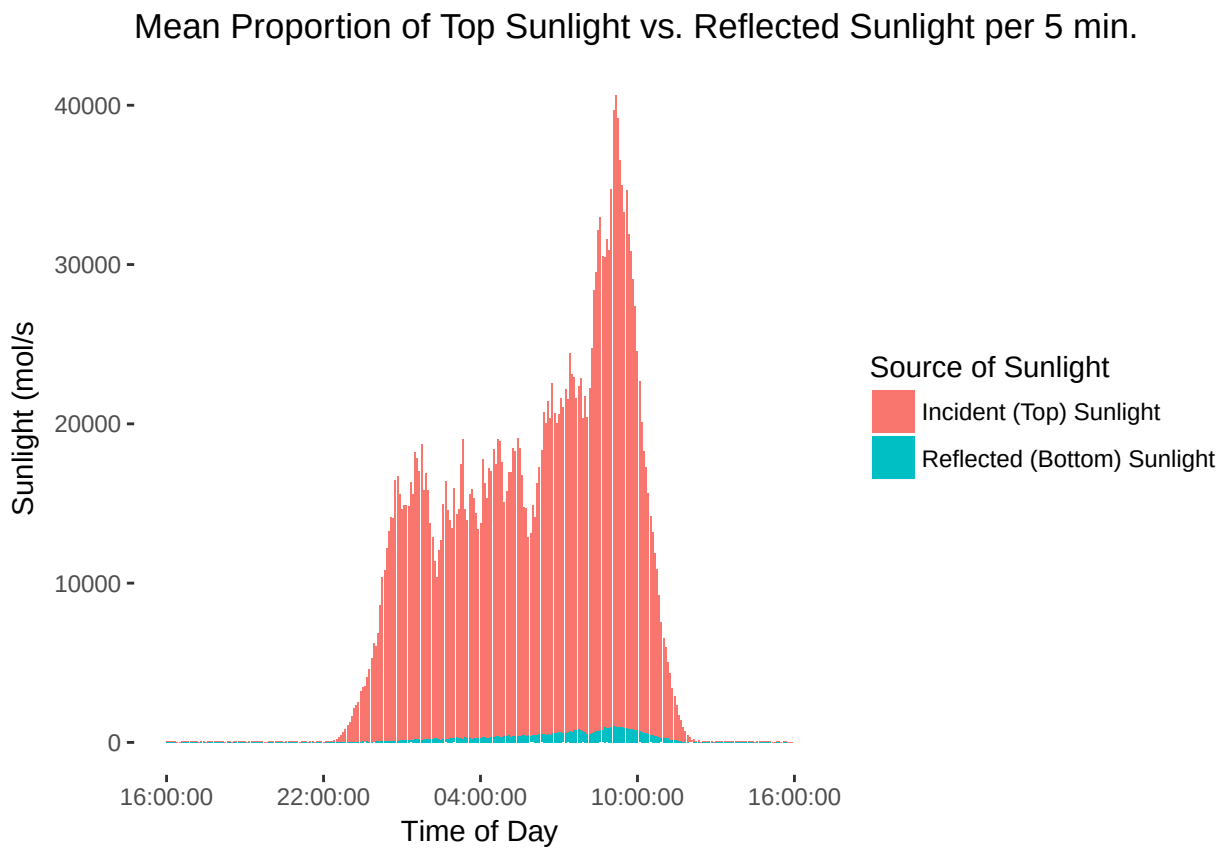



## 4.3 Third finding

My third finding relates the difference between the amount of sunlight that the top and bottom of the tree received on average, across the course of the day. The data indicate that the tree does not get much light at all during the night (whatever it does get may be coming from the moon), and the top of the tree gets the lion's share of the light. Keeping in mind the likely bias resulting from the premature deaths of lower-placed sensors, these data support the notion that the top of the tree is brighter, whereas the forest floor is much darker for most of the day.

My motivation for making this plot was to recreate the last row of Figure 4 in the paper, after discarding the outliers that I found strange. Moreover, whereas the authors only plotted data for one day (May 1st), I was interested in seeing what the microclimates looked like in the overall study. Faceting these by day would also be interesting for seeing seasonal changes, but I omit that analysis due to space constraints. Regardless, I think incorporating the full data yields better estimates, as any one day is not necessarily representative of the system. In fact, the authors picked May 1st *because* it had abnormal weather fluctuations - which in my

mind, makes it a poor candidate for inference.




# 5 Discussion

Overall, my main finding was that I have serious doubts about the validity of the data collection instrument. While the authors had interesting findings regarding the microclimates surrounding the redwood tree, they failed to account for the fact that their instrument was prone to be affected by the very things that it was measuring. This two-way causal relationship makes their inferences hard to believe. While it is certainly possible that the introduced bias is minimal, there is no way of knowing for sure without replication.

The main issue here is that the authors did not contemplate that height and distance from the center of the tree trunk would be related to battery failure. My measure for battery failure (dying before the end of the study) may not be perfect, but it is striking how many batteries actually failed. Throwing away data should generally be done judiciously, and I fear the investigators were too quick to discard large amounts of data based on voltage drops without investigating what caused voltage to drop. The relationship between height, node placement, humidity, voltage drop, and battery failure (and consequently extreme measurements) is convoluted, but tells a compelling story about why so many nodes failed to collect data.

The most alarming aspect of this trend is the difference between interior and edge nodes. Every edge node was below the 60m mark, and not coincidentially, all of them died early on in the study. Interior nodes that were lower down on the tree also seemed to die more. It is likely the case that humidity has a strong effect on battery life, and edge nodes were not well protected from high humidity. Interior nodes that were lower down the tree may have fared a little better than equivalent edge nodes, but ultimately largely succumbed to the elements as well. Interior nodes that were above the 60m mark (and therefore in warm, dry weather) had a much better chance of surviving, which makes it very difficult to assume that the quality of data collected by each node was roughly equal, regardless of where the node was placed.

Moreover, the way these nodes were networked suggests that they are likely to be fundamentally different. Edge nodes were largely networked to other edge nodes, and the same was true for interior nodes. In my mind, this indicates that a general systemic failure in either network was likely to affect most of its members. This was easily the seen in the case of the edge nodes, which I suspect all failed early because of their exposure to humidity and distance away from the trunk of the tree. While I cannot observe all of the relevant covariates, the network analysis helped illuminate that the interior/edge distinction was not trivial in terms of predicting device health. Again, I suspect the difference is largely a function of exposure to the elements, but there could be another explanation like a manufacturing defect in all of the edge nodes.

Essentially, my proposed narrative is that sunlight affects things like temperature and humidity, but those things also affect battery health. While the investigators were primarily interested in the relationship between sunlight, temperature, and humidity, and my analysis shares the same intuition, I do not believe they adequately accounted for those conditions tampering with the quality of their instruments. There may be other causal narratives that better explain battery failure, but the overall case remains that failure was not random. My narrative is largely based on the available data; I do not have insight into whether things like rain, wildlife, etc. affected battery failure, but these are certainly plausible.

In terms of limitations, I want to flag a few. First, I did not thoroughly examine the difference between the "log" and "network" sources of data, and these could have told interesting stories, or explained some of the phenomena that I am seeing. I also base my models based on my own guess at what the strange voltage readings meant - if this was wrong, the effect I am seeing exists but is exaggerated. Finally, I did not spend much time investigating nodes that reported two different voltage readings (presumably because they were keeping local and network logs), so my estimates may be biased in that I overestimate battery death. That being said, I do not suspect that accounting for these factors more thoroughly would substantially change my central insight that the measurement instruments' efficacy correlated with their position on the tree.

# 6 Conclusion

To conclude, the major lesson I learned here was that statistical analysis yielded insights into the mechanical workings of the study that were not clear from the paper itself. There was a rich relationship between where a node is placed (both vertically and horizontally), the sunlight it receives, the temperature/humidity it experiences, and whether its batteries maintain an adequate charge. The authors were excited about the potential for remote sensors to assist with macroscopy, and the most significant conclusion here was that there is still more work to do. They identified a clear problem with the difficulty of measuring microclimates before, and it seems to me that the remote sensors could still be improved to solve that problem and generate more accurate readings.

# 7 Bibliography