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

Alexander J. Brandt

September 14, 2017

1 Introduction

In this report we will investigate the data cleaning and analysis methodologies undertaken by Tolle et. al. in their study of climate dynamics as experienced by a California coastal redwood.

The data is recieved without any explanation as to the variables or their collection methodologies except for those that are specified in the paper, and imply questions surrounding their collection and interpretation. The redwood sensors (or "motes") originally were intended to pass data to a central computer via wireless network. It was later determined that the wireless data collection system was faulty or defective, and so as a result, digital logs stored on flash memory chips within the sensors themselves proved to be the most reliable source of data according to the collaborators. Both the networked collected and log collected data sets were made avalible.

2 The Data

2.1 Data Collection

The sensors were built with the following sensors, though the significant ones are bolded.

- **Time – each sensor tracks the time of each measurement taken. It accomplishes this by briefly turning on once every 5 minutes to both ensure periodicity and conserve battery.**
- **Temperature – a standard electronic thermometer, probably based on resistors with a known thermal drift**
- **Humidity – a measure of the amount of water vapor in the local atmosphere surrounding the mote. Given as a percent humidity rather than mmHg or atm.**
- Barometric pressure – A barometer measuring atmospheric pressure. Not used to do sensitivity issues.
- **Light Levels (Photosynthetic Active Radiation) – a photometer measuring photosynthetic active radiation in PPFD (photosynthetic photon flux density). Values were taken at the top and bottom of the mote.**
- Light Levels (TSR) – a photometer measuring total active radiation in a broader spectrum. Not used do to sensitivity issues.
- **Voltage – the remaining voltage in the mote's battery.**
- **Node ID – the numerical identity of the node that has collected the data (important for associating physical properties later).**

2.2 Data Cleaning

Since the redwood data set was collected from two sources (the network and the backup data logs stored in the sensors), the concatenated data set might be redundant. We attempted to check the consistency of the two constituient data sets before blindly proceeding with the log based data set as the authors suggested.

1. The authors of the paper suggest that the log data set will be more complete and therefore useful than the network data set. We inner join the two as data frames and check to see that their values correlated correctly. All measurements but voltage are well correlated. Voltage comparisons between the two are problematic. This will flag voltage as an important variable to consider in the data cleaning.

Measurement	Log/Net Correlation
Humidity	1
Temperature	1
Incoming Light	1
Outgoing Light	0.9999496
Voltage	-0.9961888

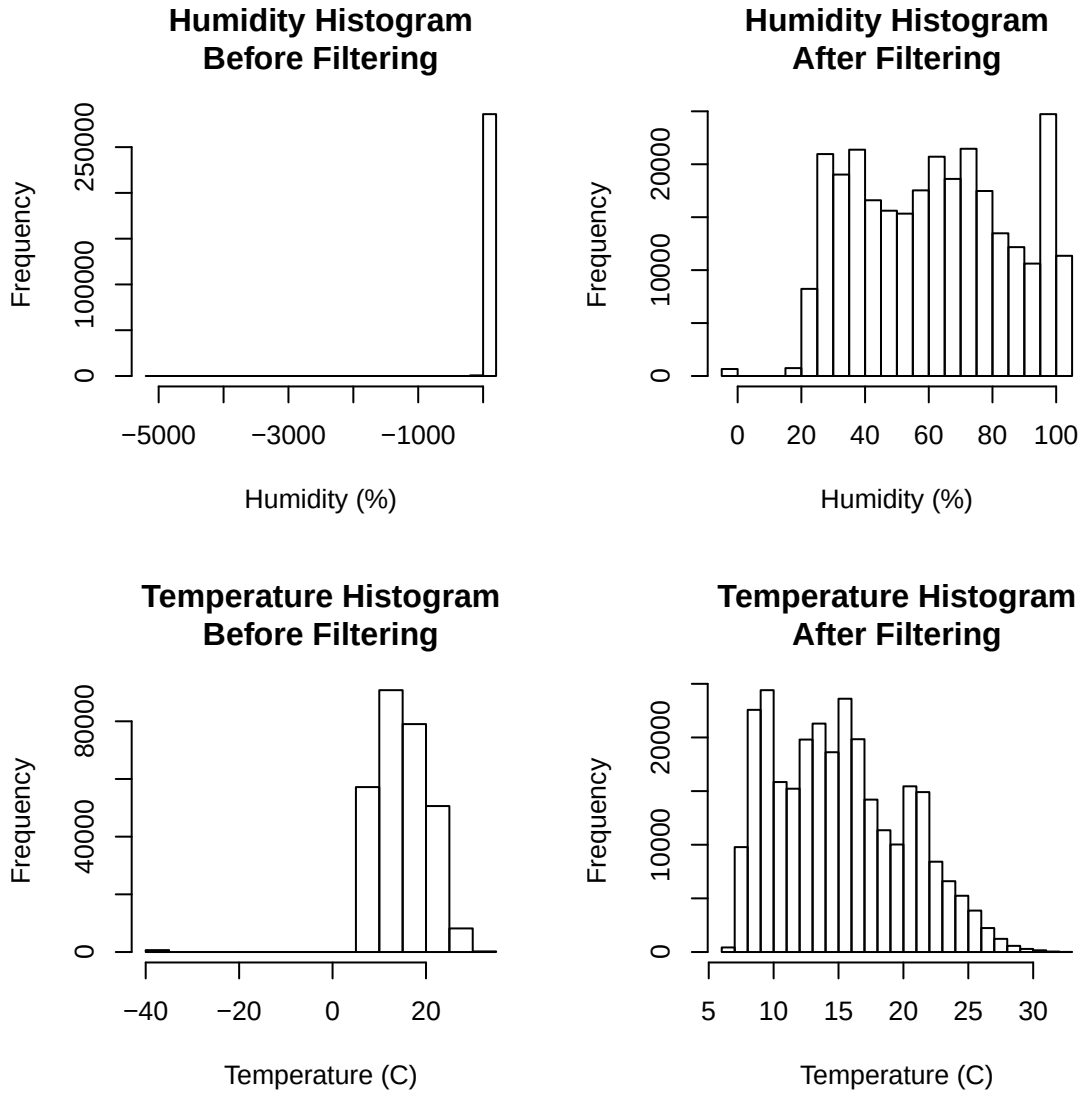
   Because the log data set is more complete, and the mutual entries appear to be mostly consistent, we proceed with using the log data set.

2. We notice a bulk of the voltage exist at exactly 0.580567 V (26040 entries), which is probably a default value set by the firmware. With such a large volume of the observations corresponding to this value, we leave them in for later investigation (this will form the basis of our first observation).

3. We notice the time in the log data set is systematically wrong (it is a constant value, probably from when the researchers downloaded the data from the motes). We perform a merger with a table that corresponds between "epoch" (the serial counting of 5 minute intervals), and the local time in UTC. We then convert the UTC to Pacific Time so we can have an intuitive understanding of daylight.

4. We remove extreme outlines from the data set for temperature and humidity. For our criteria, we allow values slightly outside physical reality (slightly less than 0% and slightly more than 100%, since this is within the error of the instrument). But we remove humidities of less than -100% or more than 200%. We also remove temperatures of less than -10 C (since there is a large, discontinuous cluster at -38.4 C, and this value is simply nonsensical given the local climate).

5. Finally, we selection just the nodes that correspond to the interior tree (there are two trees in the data set), since the paper only seemed to analyze the interior tree, as well as nodes that fall within the 0 - 200 node id range (there was one node with an invalid ID).



2.3 Data Exploration

3 Graphical Critique

3.1 Figure 3

This figure has several issues. I found the first set of histograms to be relatively useful, especially with respect to troubleshooting my own analyses.

The temperature and relative humidity are reasonably cogent. The incident PAR and reflected PAR graphs though have very little value, given that they show what seems to be a constant daily trend. If there is any other trend it is incredibly difficult to ascertain.

Given the distributions shown in part (d) of this figure, these seem fairly redundant. The boundary between the upper quartile and the outliers are very faint, almost to the point of illegibility in the case of the incident PAR and reflected PAR. It may have been better to remove these figures in order to make part (d)

larger. Or maybe represented as a heat map with color corresponding to probability density.

3.2 Figure 4

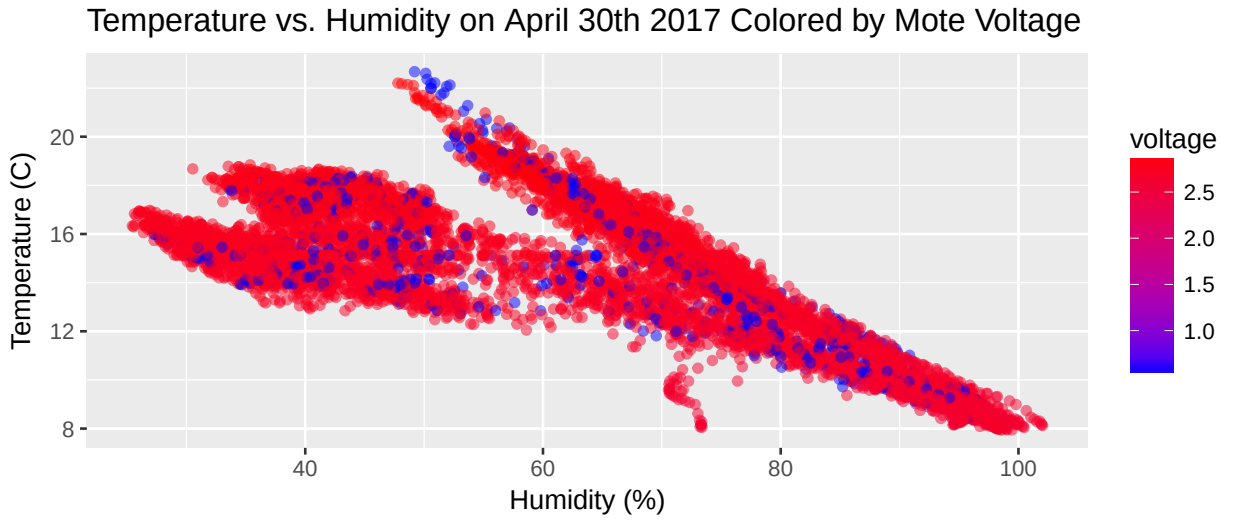
For the first two time trajectory graphs, I think the colored lines don't add very much to the comprehension of these graphs. Probably better would be an average trajectory with a standard deviation, and min/max lines. This would allow for the appreciation of the variance of the measurements at any given point in time, while making the figure feel less cluttered. The summary figures seem fine, except for the Reflected PAR graph which is very hard to read and whose mostly empty plot seems a waste of space. There coloring of the motes points (blue/pink) is never explained.

4 Findings

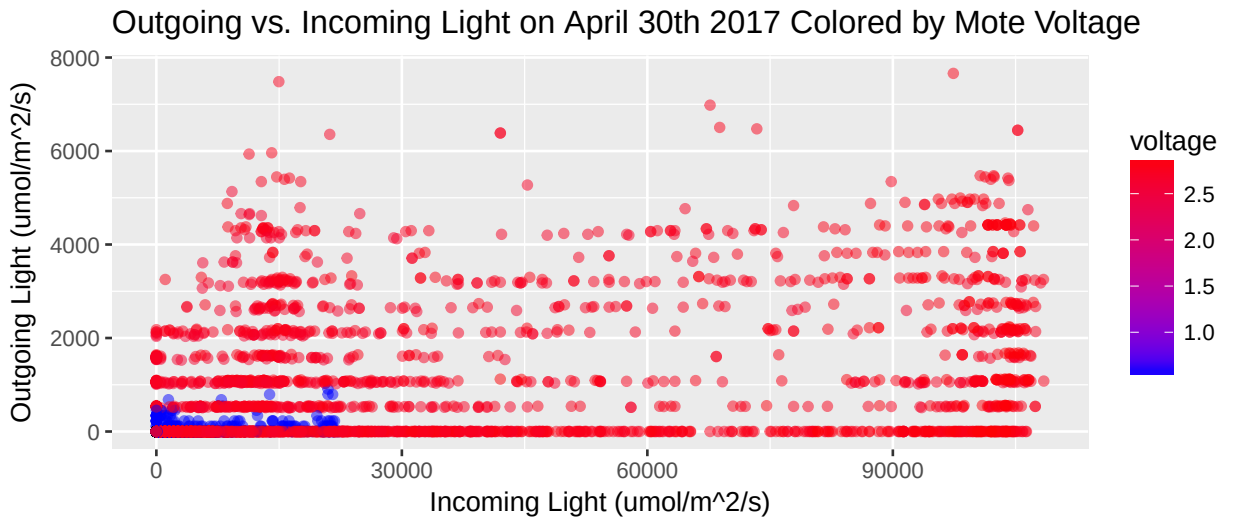
To restrict our analysis we pick April 30th (one day before the authors' chosen day) in order to select a day early in the experiment, where a large number of intact/productive mote recordings can be found, and to show some different conclusions from those drawn by figure 3 and figure 4.

4.1 First finding

When we cleaned our data set we assumed that the large spike in voltage at about .58 corresponded to a systematic failure of the voltage analysis but we will now rigorously investigate that claim. Let us take two of the most temperature and humidity, and analyze them in a pairwise graph colored by the voltage of their relative mote. Here we see that the measurements at low voltages are well blended with the measurements at the higher voltages. Indeed, we see no relative difference in the patterns, which we have plotted at a low alpha value so as not to obscure the general trend. This graph can add support to our revised data cleaning practice of including measurements from sensors that register with a voltage of lower than 2 V (or even 1 V) to add a large number of humidity and temperature readings. But what about for the other units were are interested in? How would low voltage affect light flux at the top and bottom of the mote?



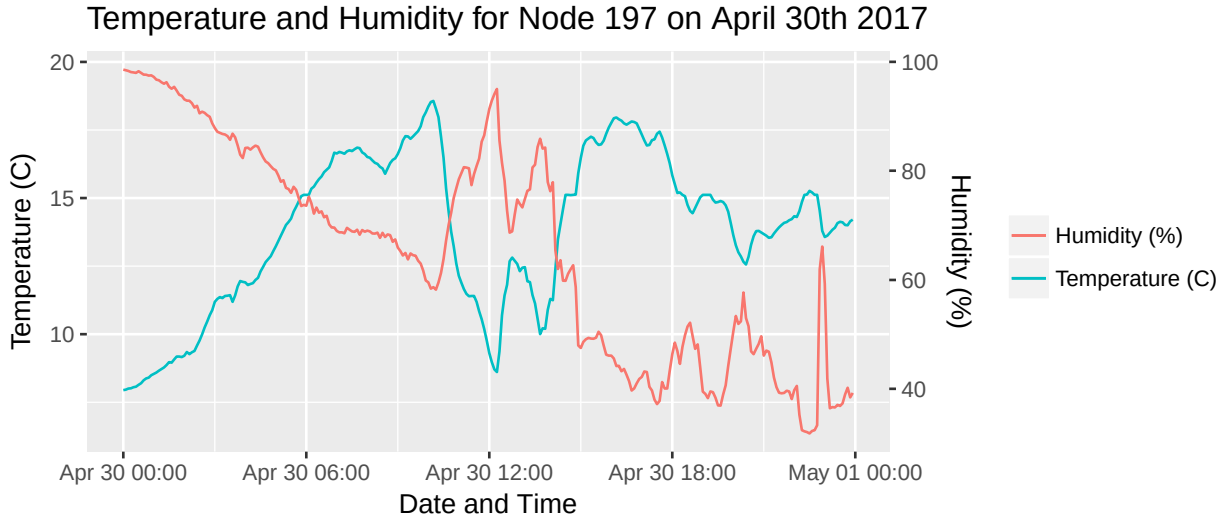
As opposed to the humidity and temperature readings, the mote voltage does seem to be prohibitive for understanding photometer readings. The low voltage motes are much less senstive to PAR sunlight than the adequatley charged nodes. We could hypothesize that this could be because the photometer elements in the motes are more power intensive than the humidity or temperature sensor. Regardless, we should be skeptical of our light readings from motes with low battery voltage.



4.2 Second finding

Again, given that we are working with the data without any correspondence with the authors of the paper, we hope to recapitulate basic physical principles as a check on our data clarity and cleaning principles. Here we examine basic meteorological relationships on April 30th.

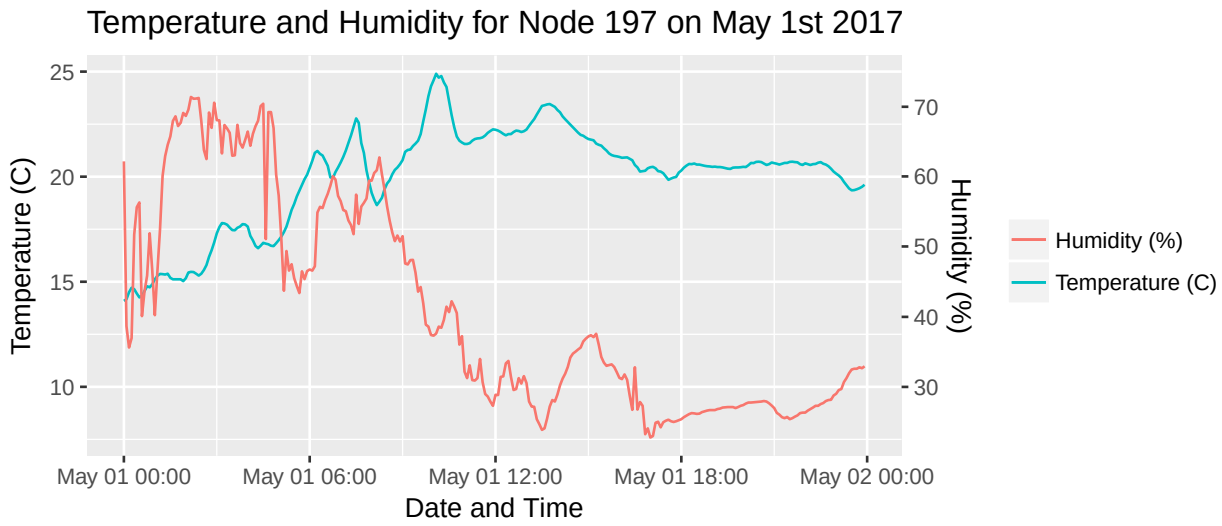
Basic chemical and physical principles tell us that warmer air can bind more water than cooler air (this is one reason for the "haze" seen in cities on muggy days). So, assuming a constant volume of water, an increase in temperature usually corresponds to a decrease in relative humidity. This relationship is recapitulated in this graph (and has a not unreasonable correlation coefficient of -0.8849986).



So when the authors seem to suggest that on May 1st the tree experienced a microclimate that didn't obey conventional wisdom around humididty/temperature relationships, I became suspicious and re-plotted it with my expanded data set that included the lower voltage motes (given that I suspect their climate data is fine to use for the analysis for reasons stated above). Indeed, the conventional relationship between humidity and temperature as a function of time seems to be re-established. I believe their conclusions were potentially faulty because they overzealously cleaned their data.

I would need to check for the trajectories in aggregate, rather than just this one node, to say definitively if

their conclusions were faulty, but in the interests of looking at different types than just the ones presented in this paper I did not do so.

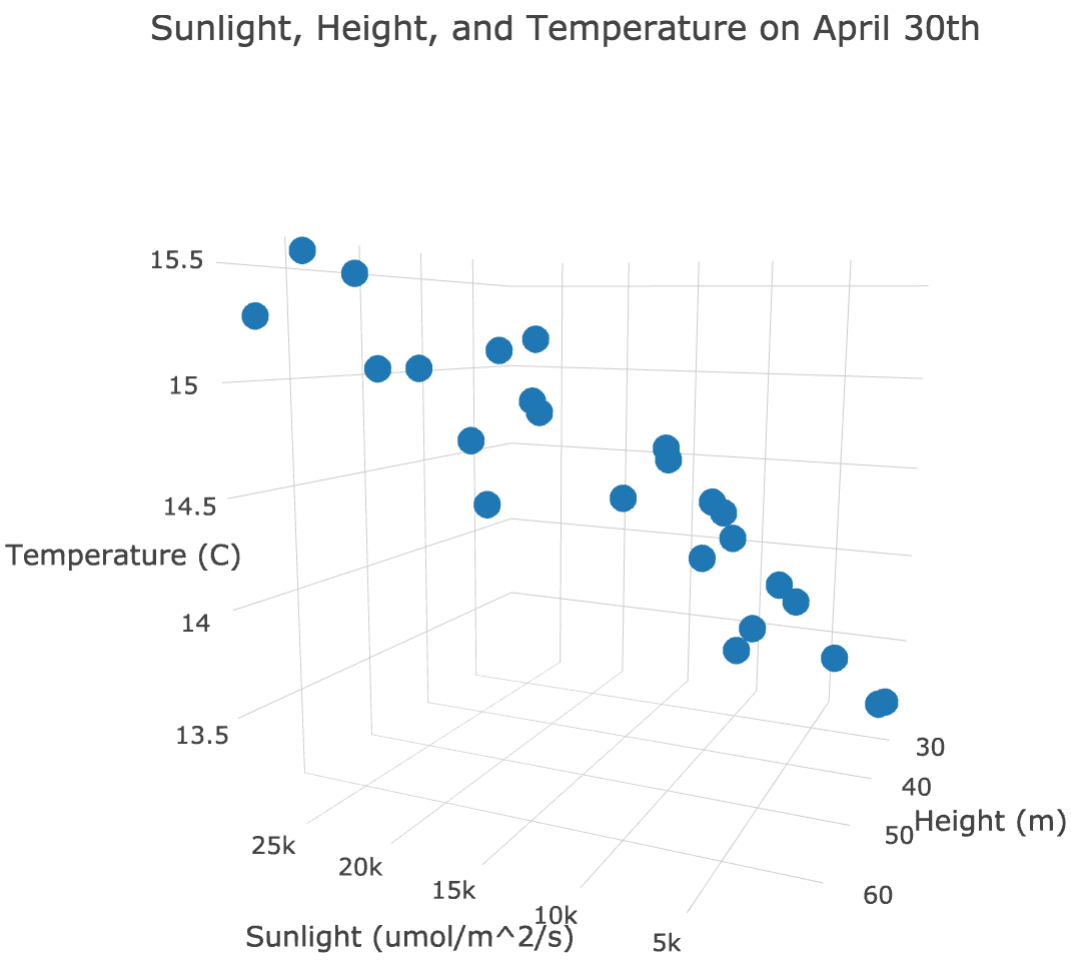


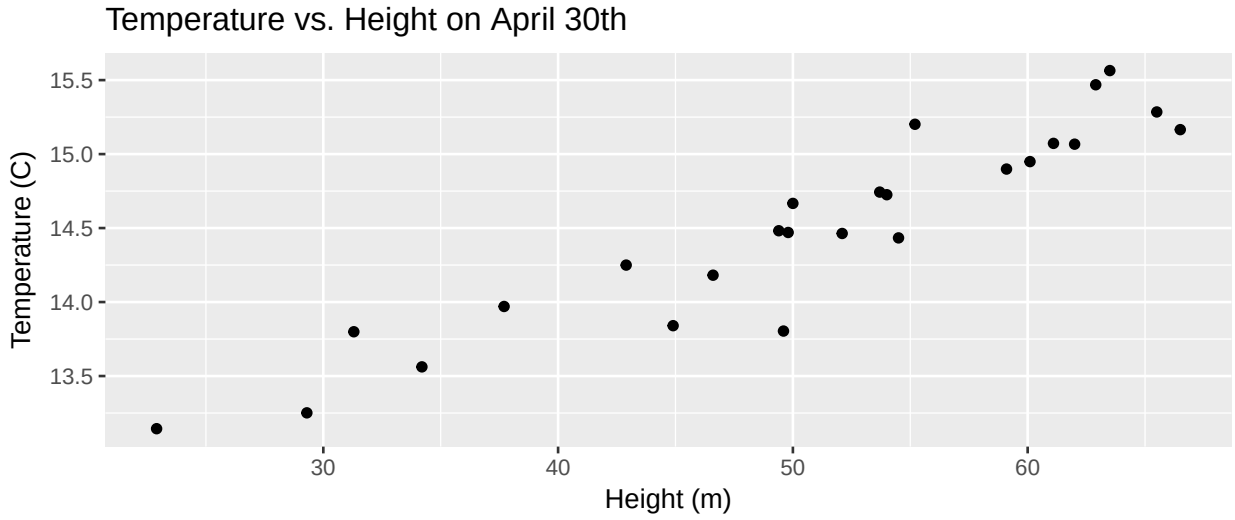
As an aside, with these two trajectories of information, we could calculate the dew point at each sensor for the day, if we cared to.

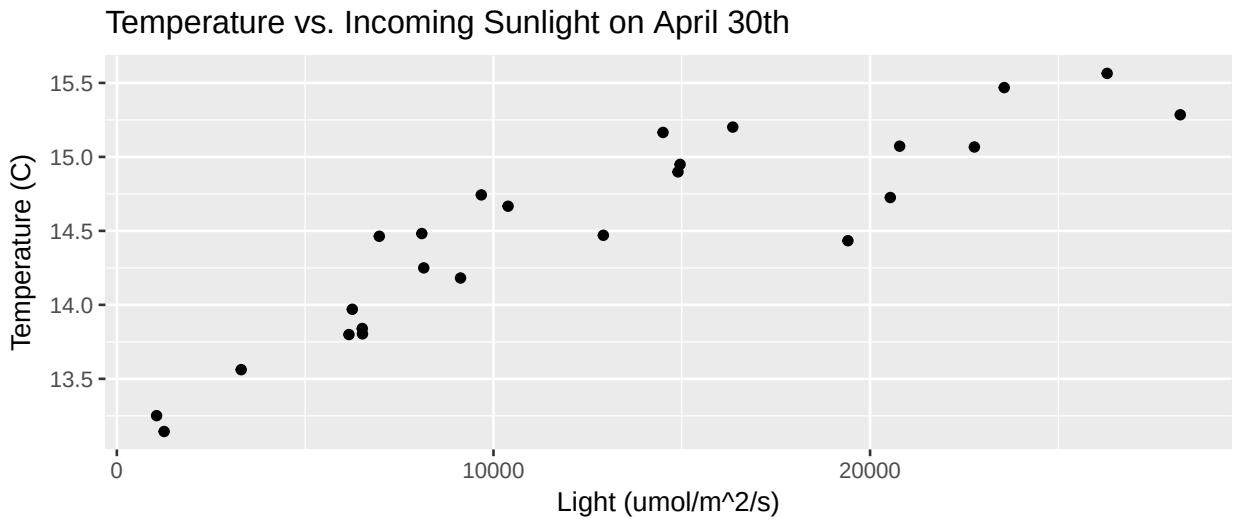
4.3 Third finding

Another consistency we would like to investigate is the relationship between temperature and sun exposure. Redwoods are famously large/tall trees. In order to isolate minimum shading from the leaves and branches, we will investigate the relationship between node height and sun exposure. We first attempt to check our assumption that tree node height corresponds with average daily sun exposure (which we will determine by looking at values from the top photometer of the mote). We will need to subselect for voltages above a certain threshold (here, 2 V). This filtering only removes 3 nodes from the data set. A screenshot from the 3D plot is shown below, but the interactive graph can be found in this directory (filename: graph3.html).

By moving around the interaction you can visualize a coherent trend between the three variables. Cross sections of the graph are also shown below.


Sunlight, Height, and Temperature on April 30th
15.5
15
14.5
Temperature (C)
14
13.5
25k
20k
15k
10k
5k
Sunlight (umol/m^2/s)
30
40
50
60
Height (m)


Temperature vs. Height on April 30th
15.5
15.0
14.5
14.0
13.5
Temperature (C)
30
40
50
60
Height (m)



As we might expect, increased height and sunlight exposure correspond to higher temperatures.

5 Discussion

The data size was not challenging computationally, but it was necessary to subset the data by days in order to show the trends and analyses clearly. Further work could be done representing the month's worth of data effectively all at once.

6 Conclusion

This lab offered an opportunity to see how different data cleaning methodologies alter the conclusions of an experiment. It allows showed how choices of which data to use can vary between question to question of the data set. Through the redwood meteorlogical and sunlight information, we are able to establish interesting physical relationships that resonate well with our understanding of basic physical and biological principles.

7 Acknowledgements

The author gratefully thanks Max Andrew Gardner, a fellow student in this course, for assistance with the lab.