Lab 1 - Redwood Data, Stat 215A, Fall 2020

September 17, 2020

1 Introduction

The premise of this investigation begins with the question of what are the variations in different climate metrics in an area localized to a specific redwood tree: i.e. the microclimate of a tree. The interest is primarily a result of the known fact that the weather around a specific tree has fairly substantial temporal variations in factors such as temperature and humidity. Various biological factors and processes of the redwood tree contribute to these temporal changes; an example is transpiration where water moves from the roots of the tree upwards toward the leaves where they evaporate, affecting the humidity. The goal is to better understand to what extent does the redwood tree itself affects this climatic temporal fluctuation independently of the existing variations in weather.

2 Data

Tolle et al. [2005] placed a number of different (a total of 80) remote sensors, called "motes", at various positions relative to a specific redwood tree. The positions differ in their height, their distance from the tree, and their cardinal orientation. The motes collect a variety of differing measurements regarding the climate at the time every five minutes within the timeframe. The differing locations of the remotes sensors and the climatic measurements made frequently capture both the physical and temporal variations of the microclimate specific to the redwood tree and provides the data to begin analyzing the effects of the redwood tree on the climate.

2.1 Data Collection

The motes measured every 5 minutes the temperature, the relative humidity, voltage readings on the battery at the time of measurement, as well as the amount of direct sunlight and reflected sunlight (measured through the bottom of the remote). The temperature is measure in degrees Celsius; the relative humidity describes the percentage of water vapor in the air relative to the amount of water that the air can support. Both sunlight variables are measured through photodiodes recording radiation. The node id is a unique identifier of which remote the recording is from, and the epoch is an identifier of the 5 minute time period from which the measurements were supposed to be taken.

2.2 Data Cleaning

The raw dataset received had quite a few inconsistencies that had to be addressed: the following plots highlights some of the issues that had to be addressed.

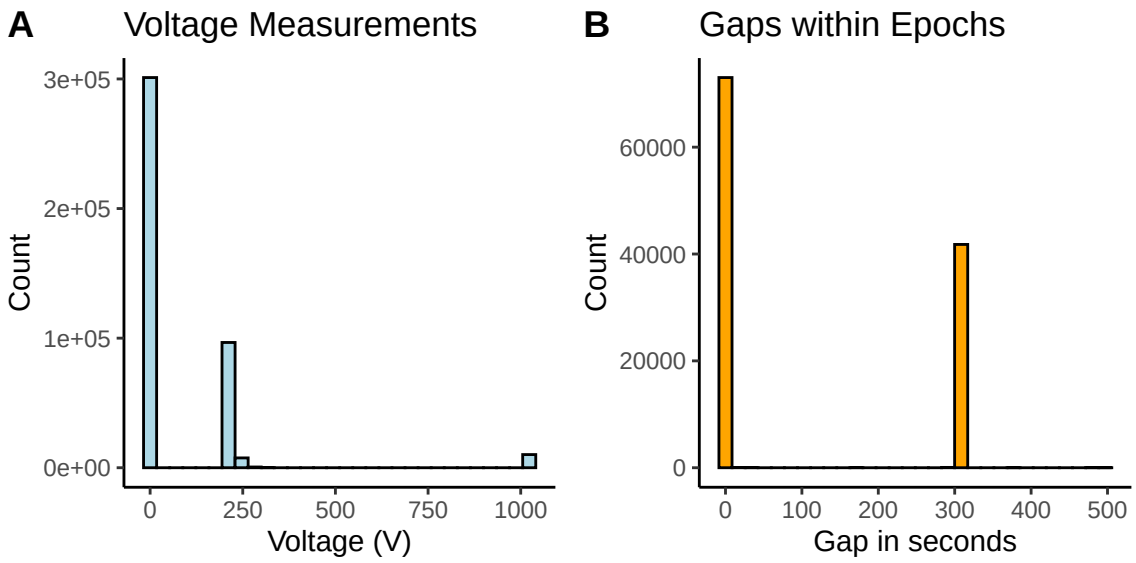

Figure 1: Issues with Voltage and Time Measurements

Figure 1(A) highlights some of the issues found within the voltage measurements: there is quite a drastic range between with most observations falling between 0 and 5 volts, some in the around of 200-300 volts, and several more over 1000 volts. Note that the battery of the device for the remote only supports voltages between 0 and 5 volts, so there are clearly issues for the other ranges. Figure 1(B) indicates issues in how the time data is recorded: the epoch is used to group measurements from the different motes together into being roughly at the same time every 5 minutes. Figure 1(B) is computed through grouping the data by epochs and computing the greatest difference between recorded times within the groups. Most gaps in recorded time are within an acceptable interval i.e. 10 seconds, but for certain epochs the gap is more than five minutes which is longer than the periods in between measurements. This also needs to be addressed in cleaning the data. Figure 2 highlights the final issue with the data: many data points had a date that fell outside of the specified time interval of collecting, primarily in November where the data recording dates were between April and June roughly.

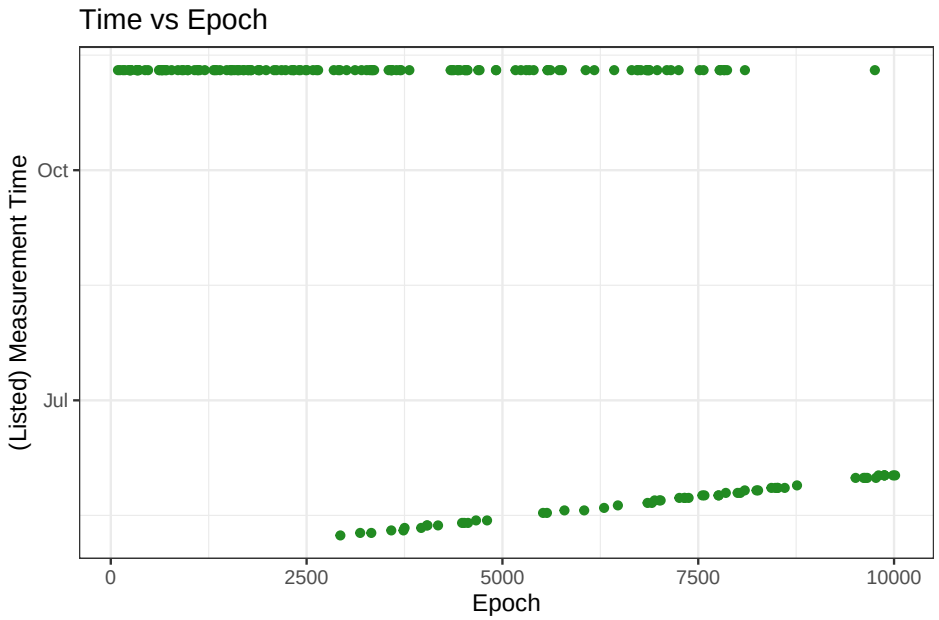

Figure 2: Issues with Recorded Dates

Upon further inspection, it became apparent that some of the data points in November might have been repeats of the measurements made during the summer since when some data points are matched via epoch and node id (which are supposed to be unique identifiers of data points) the measurements are identical with the exception of the time and voltage. Inner joining on node id and epoch and then selecting the datapoints that have differing times gives 209496 data points; inner joining on everything except for time and voltage and then filtering for points that have differing times gives 209164 so most data points in November which match up in epoch and node id with data in the time of the experiment are just the same data points. It is also interesting to note that by inspection it seems most of the out of range voltage measurements of 200+ are in the summer measurements but the reasonable voltages are found in the November measurements.

The cleaning procedure then proceeded as follows: first the data was inner joined on itself matching every column except for the voltage and time of measurement and then the data points where the times of measurement differed across the join were selected for. Then the measurement of voltages corresponding to November were kept and the ones for summer where dropped due to the numerical discrepancy. The justification of these steps is that upon cleaning the original dataset and filtering first out for dates within the range of recording and then for reasonable voltages, the end result was no data points: performing the join would provide data points with accurate voltage measurements with the risk of dropping certain points within the time frame that did not have a match from the data points in November. Upon dropping the NAs, most of the issues regarding the gap within an epoch was resolved with the largest gap being roughly 3 minutes, but most of the gaps are under 10 seconds. After these steps there were still some points that were outside of the accurate range, so points who's corrected dates that were still outside the time frame were removed, as well as those with high voltages (i.e. > 190). The result of this cleaning procedure was a dataframe with 72,000 data points and then joined with the lookup table on dates for convenience.

2.3 Data Exploration and Reality Check

Next we proceed to perform a basic sanity check on the data and understand some of what's going on in it. Figure 3[1] highlights the average humidity from all the nodes across every day and mote as well as the average

[1]apologies about it saying September, converting the time of day to a datetime object automatically as today as the day and I couldn't figure out how to get rid of it

temperature. The graph reflects nature: the average temperature is higher during the day and highest in the afternoon and lowest at night, and the average humidity is highest at dawn and lowest during the day.

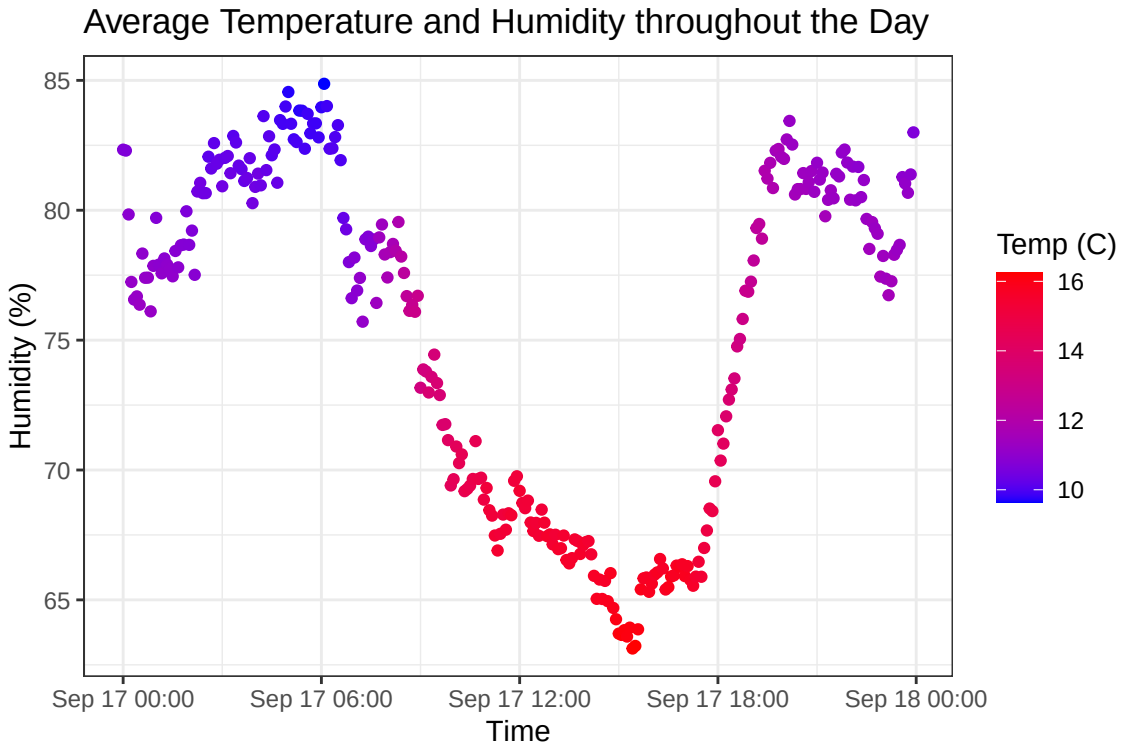

Figure 3: Reality Check

Figure 4 measures the average sunlight per time throughout the day across all the remotes of both the top photodiode and the bottom one. The plot roughly represents what one would expect: the solar radiation is low at night, increases in the morning, stays high throughout the day, decreases in the evening and returns. Additionally as expected the sunlight from the bottom photodiode is consistently lower than the top one. Something that may be surprising is the measurement at night: there is still non-zero measurements of sunlight despite common sense, however that depends on the actual measurement itself. Additionally it seems that there is most sunlight in the afternoon whereas one might expect it to be the highest at noon. However the general shape of the trend reflects what one would expect in reality.

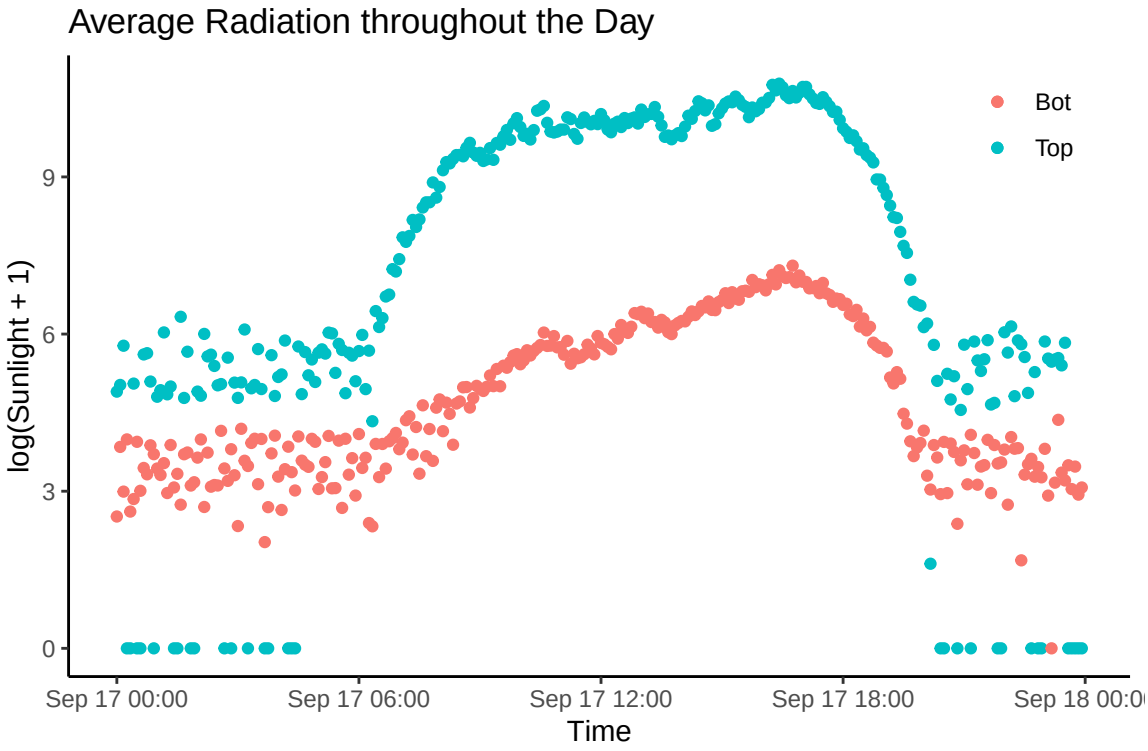

Figure 4: Sunlight Reality Check

3 Graphical Critique

- Critique the plots in Figures 3 & 4 in Tolle et al.
- What questions did they try to answer? Did they answer them successfully?
- Did they raise any questions not addressed in the text?
- Would you change them at all?

Figure 3 is a bit complicated and on the smaller side, making it somewhat difficult to read but given enough effort and dedication its possible to make some sense of it. The purpose of the plots however is not to convey any sort of deep meaning but to demonstrate how one might go about analyzing multidimensional data indexed by two set of variables (time and height) alongside the value of interest. To improve Figure 3, it would be reasonable to only include 2 of the values, rather than all 4. The purpose is not to convey some meaning about individual values but how to illustrate data so plots for all 4 measured variables are not required. However Figure 4 is an unmitigated disaster: the plots on the left are overall too complicated and cluttered to make any reasonable sense out of it, even after staring and zooming in for quite a while. The plots would be better of using a subsample of the motes rather than all of them to make the plot easier to read and understand, as well as easier to label with a legend.

4 Findings

The objective of this data is to study how the biological processes of the redwood tree affect the microclimate around it: unfortunately the data itself does not provide any direct measurements of this but we can use various data as a signal for it. The measurements of sunlight could provide a rough approximation for how much photosynthesis and transpiration goes on: the justification is that photosynthesis is directly related to sunlight and transpiration occurs via the stomata opening during photosynthesis to release carbon dioxide

and water as well, affecting the humidity. The next few findings relate primarily to how sunlight is related to the microclimate: using sunlight of the top photodiode as a proxy for photosynthetic activity (although it is compared relative to the reflected sunlight measured from the photodiode to attempt to control for the impact that the sunlight itself has on the weather). For clarity: we define the "peak" of the tree to refer to motes near the top of the tree, mostly 60m or higher above the ground and "canopy" to refer to motes closer to the ground, mostly 20m or lower above ground.

4.1 First finding

We begin by considering the variation in the solar radiation received by the top and bottom photodiodes for the canopy and peak layer: we take the motes in the canopy layer and average out the measurements of the top and bottom photodiodes across epochs and do the same for the peak layer. The measurements where the average of the top photodiode which are 0 are filtered out as an approximation to try and remove datapoints in which there is no sunlight. The kernel density plot in both cases is plotted below in Figure 5. Figure 5(A) is the kernel density plot of the ratio of the measurements at the canopy level, and Figure 5(B) likewise for the peak layer. We add one to each of the measurements to prevent issues with undefined values.

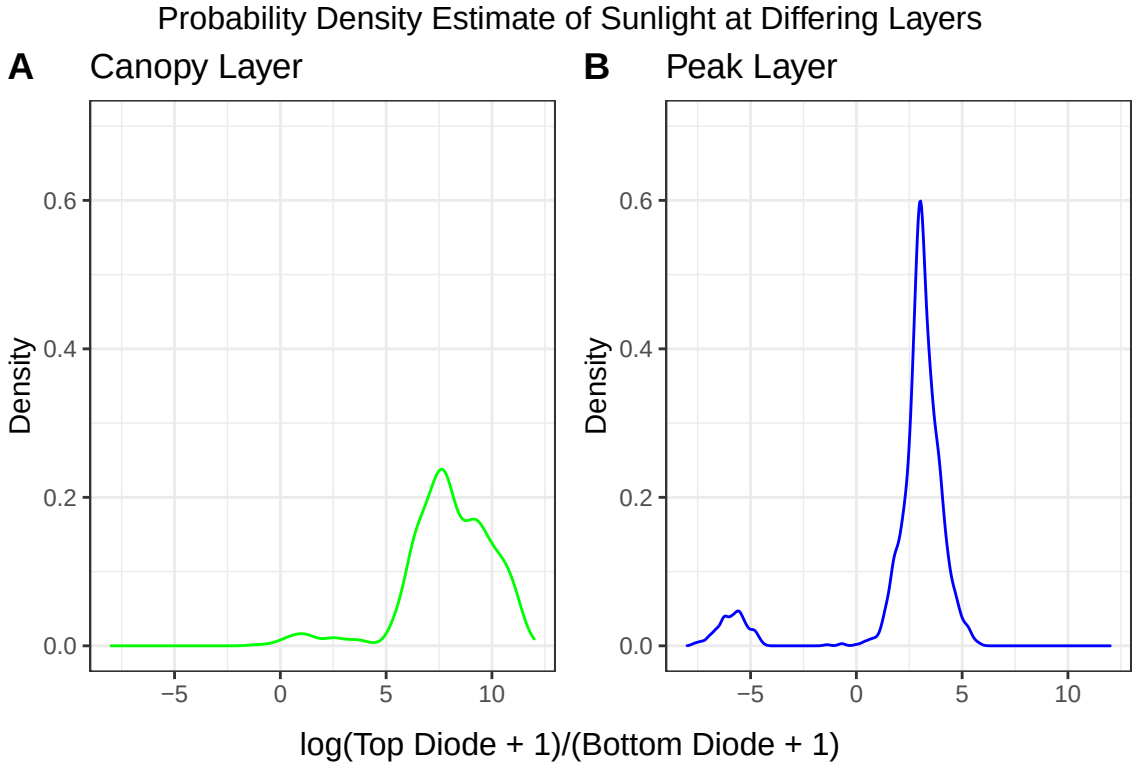

Figure 5: Variations within Sunlight Between Canopy and Peak

Note that the figure illustrates that at the peak layer the ratio of the measurements by the top and bottom diode remains consistently around 1000, whereas the ratio in the canopy layer varies between 10,000 and 1 trillion times. This is an interesting result because common sense is unclear as to what would be the expected result: at first one might surmise that they should roughly be the same because the height difference is only around 50m which is not that considerable relative to how light travels.

4.2 Second finding

Next we consider how this ratio of top photodiode to bottom photodiode at the canopy layer affects the humidity throughout the tree. We follow roughly the same procedure as above, but in addition we compute the average humidity measurement per epoch for the canopy motes and the peak motes. The results are plotted in Figure 6 below.

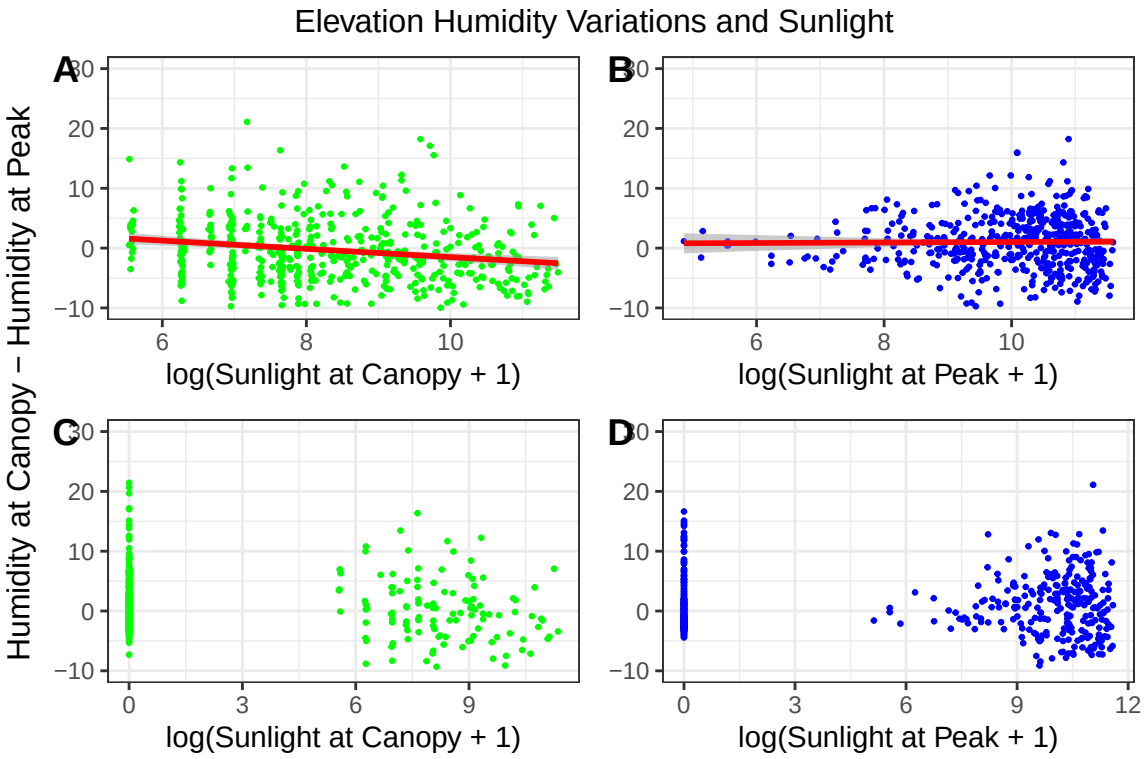

Figure 6: Variations within Sunlight Between Canopy and Peak

All the plots are scatterplots of the average humidity at the canopy minus the average humidity at the peak per epoch against the log of the sunlight at different points. Figure 6(A) and 6(B) are plots with the top photodiode measurement equal to 0 removed (again to try and control for measurements at night), whereas 6(C) and 6(D) includes such measurements. The regression line is added to figures (A) and (B) and it becomes apparent that there is a negative relationship in (A) between the amount of sunlight received at the canopy, and the humidity gap between the bottom and the top. Note usually the humidity at the peak of the tree is consistently lower than the humidity at the canopy layer. This means that the more sunlight there is received by the canopy layer, the gap between the humidity of the canopy and the peak is smaller, and that this effect is less pronounced when considering the sunlight received at the peak layer. A proposed explanation for this phenomena could be the increased sunlight means that there is more photosynthesis occuring and thus transpiration and since water vapor rises it accumulates throughout the tree at the peak.

4.3 Third finding

Next we consider if we can capture variations between the humidity of the North side of the tree and the South side of the tree. We follow a similar procedure for the previous finding, except we compute the desired averages separately for the motes that are on Northern and Southern sides of the tree. The results are plotted in Figure 7 below.

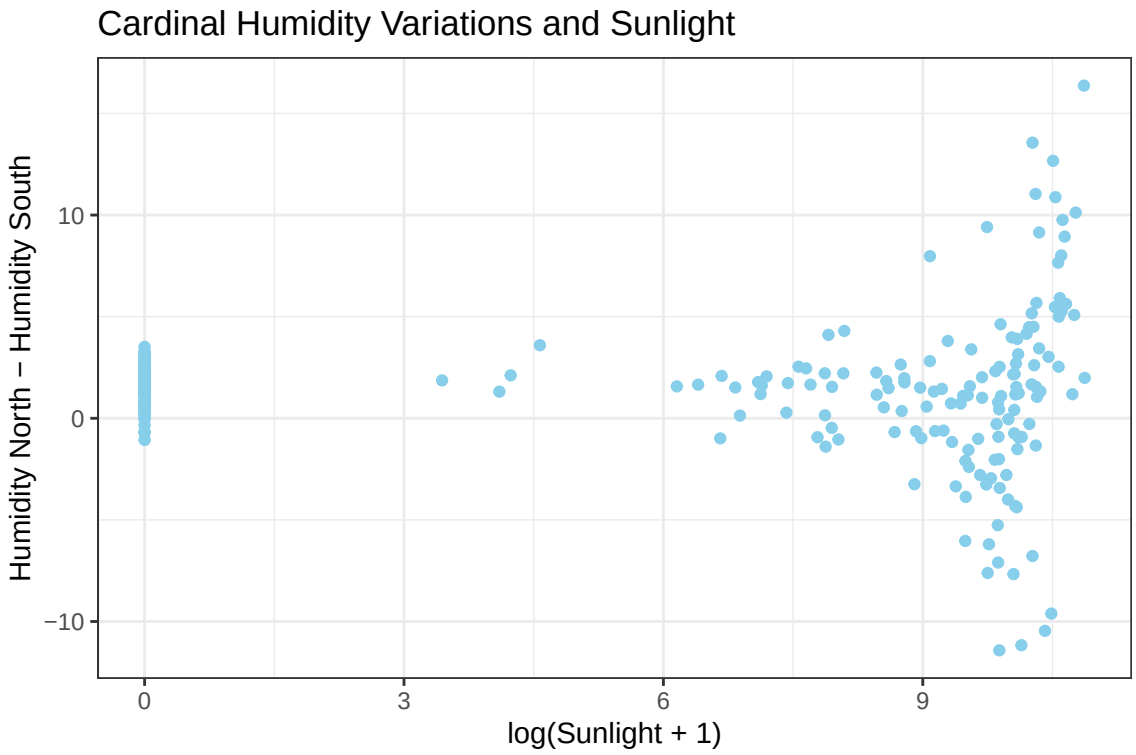

Figure 7: Variations within Humidity Between North and South

Figure 7 highlights something quite interesting: the normally the Northern side of the tree tends to be more humid than the Southern side but the difference is quite minimal at night, or at least when the measurements for the top photodiode is zero, but the variation grows quite dramatically as the sunlight increases. One might expect that the humidity would roughly remain the same across the Northern and Southern parts of the tree as they are just separated by a few meters but we actually see quite dramatic variations in the humidity, especially when the sunlight measurement is quite high. However this result is hard to believe and it might be the case of inaccurate data.

4.4 Stability Check

In the data cleaning procedure described above, in order to resolve some inconsistencies with the data and the voltage measurement, several earlier dates were dropped since it turns out they did not have a repeated copy in the data listed as November. The stability check data cleaning procedure is to ignore some of these inconsistencies, filter out all the data whose date falls outside the time frame of the experiment, and duplicate data points are removed. We re-ran finding 3 as it was the one that seemed the most unlikely to verify the stability of the result and the results are displayed in Figure 8.

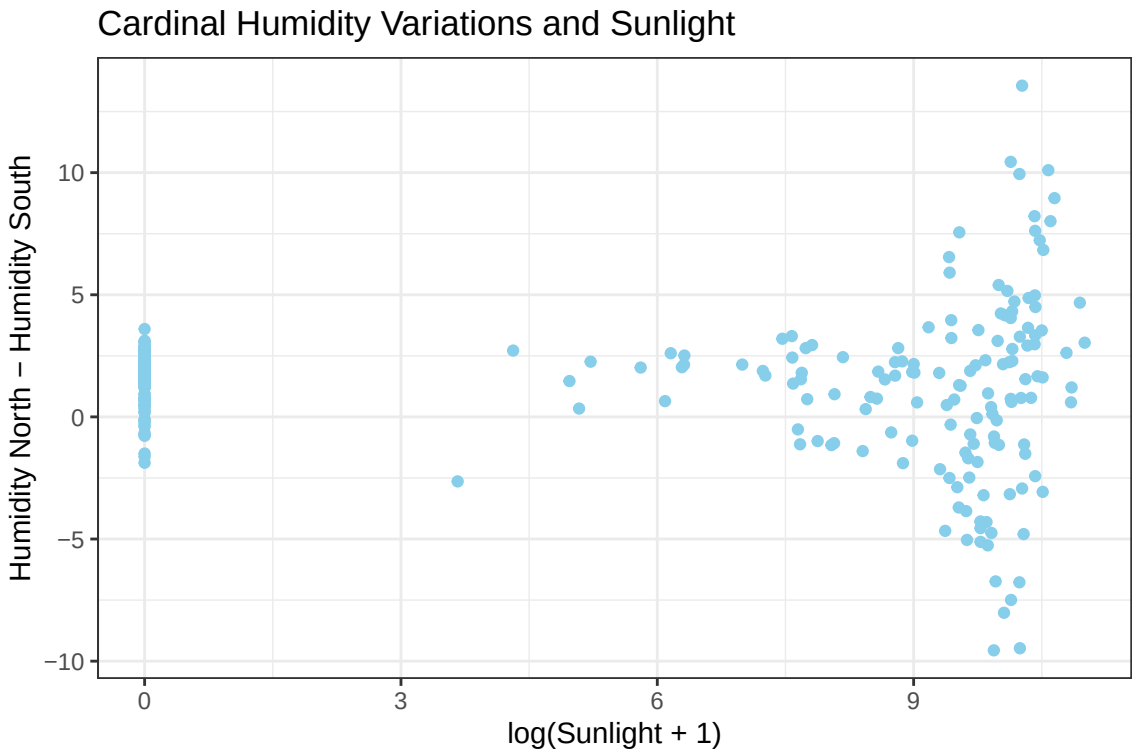

Figure 8: Stability: Variations within Humidity Between North and South

It turns out stability check says that the results from finding 3 might not be entirely inaccurate and confirms what was found there. However we should still probably dubious of drawing any sort of serious conclusion.

5 Discussion

The data set size was on the larger side and this prevented certain challenges as our assumptions and intuitions about how to clean the data are guided by manually checking individual data points, but due to the data size it always possible that there are certain kinds of issues within the data that were entirely missed in the cleaning process. The data captures the reality of the weather and solar variations around specific locations of the tree. This report is somewhat sparse on the algorithms and models: the interesting findings hint at interesting results but to fully confirm them (within the limits of our data set) we would need to either make modeling assumptions or apply certain algorithms in order to be able to draw significant conclusions. This report speaks to future data and reality by illustrating interesting findings, such as the variation in humidity between the northern and southern parts of the tree, giving evidence towards a new understanding of reality, i.e. how the tree itself affects the microclimate, and would require us to collect further data to draw meaningful conclusions. However this is all predicated on the assumptions that the data is an accurate representation of reality; otherwise all these interesting findings are fairly worthless.

6 Conclusion

In this report we've identified several climatic efforts that might have some cause in the tree's biological processes: the difference in humidity between the peak of the tree and the canopy, and the likewise the variation in humidity between the northern and southern sides of the tree. However the conclusions seem

somewhat dubious at best given the status of the data and further inquiry is required to draw meaningful, significant conclusions.

7 Academic honesty statement

In the course of completing this project, the outside help I have received is limited to discussion of certain ideas involving data cleaning and interesting findings with others. However all the code written here and most of the work done is entirely my own, with the exception of assistance from the internet. I acknowledge and understand that dishonesty involving copying other people's work without proper acknowledgement is a personal moral wrong. Additionally, as members of the scientific community we have a duty to the public to be as transparent and honest about our funding, considering the support we receive (both financially and the respect that we afforded) and that to commit academic dishonesty is to damage our credibility to the general public, no matter how small the damage.

Bibliography

Gilman Tolle et al. A macroscope in the redwoods. *SenSys*, 2005.