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

Zoe Vernon

September 14, 2017

1 Introduction

The data analyzed in this paper was collected through the wireless sensor network discussed by Tolle et. al in "A Macroscope in the Redwoods." Tolle et. al designed nodes that were placed at varying heights on two redwood trees on the California Coast. The nodes were designed to measure temperature, humidity, direct photosyntheically active solar radiation (PAR), and reflected PAR every 5 minutes over forty-four days from April 27, 2004 to June 10, 2004. The purpose of the project was to study the microclimate of a redwood tree from top to bottom. Below is a review and critique of the data collection process, data cleaning, and graphics as well as my own data cleaning, exploration, and three figures that aim to illustrate interesting aspects of the data.

2 The Data

The data was collected in Mica2Dot packages that contained sensors measuring temperature, humidity, incident PAR, and reflected PAR. The packages were designed to protect the electronics from the environmental conditions while still allowing the sensor to be exposed to the elements to ensure accurate readings. They contained two sensor boards on top and bottom in a cylindrical enclosure. The sensing board at the top of the package was used to obtain measurements of direct PAR, while the bottom board was used for the remainder of the measurements. The outside skirt of the package provided the shade necessary for the readings, as well as protection from wind and water.

In designing the packages the researchers had to consider many nuances involved in the measurements. Humidity and temperature sensing require shaded area with adequate airflow, while the direct PAR and reflected depend on sensor orientation in terms of the four cardinal directions and the differing amount of sunlight entering through holes in the canopy. Slight changes in mote orientation affect the amount of light that reaches the sensors throughout the day. Additionally, ambient PAR measurements required shaded areas and a relatively wide field of view. All of these are small details that have the potential to create inaccurate measurements.

Once the packages were designed the researchers calibrated the sensors in two phases. First, they placed the PAR sensors in direct sunlight on the roof of a building. The roof provided an unobstructed view to study the readings with varying sunlight angles. They left the sensors for two days taking readings every thirty seconds and compared to sensors with know to produce accurate results. Second, they calibrated the temperature and humidity readings in a controllable chamber, changing temperature between $5C$ an $30C$ and humidity from 20% to 90%. The packages were then placed along the tree every 2 m from 15 m to 70 m. Tolle et. al. do not mentioned how they calibrated the measurements for reflected PAR, or why they choose not to allow th chamber to reach 100%, which the most prevalent value in the forest.

In the remainder of the paper we will focus on temperature, humidity, and incident PAR. The temperature was measured in degrees Celcius, and ranged from approximately $6C$ an $32C$ throughout the study. It is know that the top of a redwood tree experiences wider variation in temperature. As sunlight heats of the top of the tree a front of warmth moves down the tree over time. Humidity, which measures the percentage of water vapor in the air, ranged from 16% to 100%. The distribution of humidity across the tree is more complicated than temperature, because fronts of humidity that move down the tree are counteracted by water being moved up from the soil. Incident PAR is radiation that has wavelength between 350 nm and 700 nm. Incident PAR provides information to researchers about the amount of solar energy available for photosynthesis and is expressed in terms of photosynthetic flux density ($\mu mol/m^2/s$).

These three variables together give a relatively comprehensive look at the important climatic factors at play in a redwood forest.

2.1 Data Collection

The data from the packages were collected and stored in a local log and transmitted over a cellular modem using the TASK system to a remote computer storage. The network was designed to wake up for 4 seconds every 5 minutes and take the sensor readings. TASK provided time synchronization for this wake up, but it is reasonable to question if four seconds was always enough time for the node to complete sensing and record the information to the log. There were a number of issues with this style of recording the data as the yield for the network had peaks at 0% and 40%, while the log had concentrations between $0\% - 20\%$ and $50\% - 80\%$. In addition to nodes dying and malfunction one reason for the poor yield of the network was that it was down at two points during the collection process. Using the network was important, because as time progressed the logs on the nodes filled up, and thus the data was only saved through the network. One potential solution would have been to recognize the size of the data that was being collected would be too large and to install bigger memory inside the nodes for the local logging.

In the event that a node failed it could be rebooted remotely. It seems reasonable that there could have been a model that was set up pre-data collection to process the results and give indication to the researchers about which nodes needed to be rebooted or required a new battery. The cellular network could have been used strategically to detect these anomalies along the way that would have then allowed the researchers to be alerted to potential issues, such as the data logs running out of storage or batteries dying. Then they would have had the opportunity remedy these issues during the study, and have more data in the end.

Of the 1.7 million data points collected only $49\%(820,700)$ were used in the final analysis. There were numerous missing values and readings that appeared to be extraneous that caused observations to be removed. The authors choose to remove all nodes that did not have voltage between 2.4 and 3, because voltages outside those ranges correlated with batteries dying, which correlated with outliers in the other measurements. Additionally, they removed sensors that did not track with other readings, and ones that failed to produce any data.

2.2 Data Cleaning

The approach that Tolle et. al used to clean their data removed many of the outliers by simply deleting readings outside of the voltage range, but also required removing many nodes that I discovered produce readings that appeared reasonable. Consequently, I choose a different approach. I began my cleaning by creating common variable names so that the data sets could be combined, changing the result time into the standard format, and deleting the observations with missing values. I choose to do so because I wanted to understand the relationship between all the variables on the tree, so I needed observations that contained all the data. I also decided to transform the voltage data for the network data, because the values were in the 200s and 300s, which is well above the voltages reported in the paper. I found a linear relationship between the log and network voltages at the same time and node and used that to transform the network voltages by

$$Voltage = \frac{442.97 - Voltage}{82.55}.$$

This produced voltages in the appropriate range that correlated with battery failings that could be seen by looking at spikes in temperature as they do in the paper.

Next I started investigating the data graphically, looking at histograms and box plots of each variable (temperature, humidity, incident PAR, and reflected PAR), as well as scatter plots of each variable over time and space. Throughout the process I zoomed in on areas that appeared to have anomalies. Once I found a group of points or single point that appeared out of place I highlighted the node (or nodes) of interest to see if they no longer trended with the other variable at that time and height on the tree. After doing so I continued to search for further outliers by looking at pairwise scatter plots between the two variables. Once I found data points that could not be explained by their position in space or time in relation to the other data I deleted those observations. I also choose to ignore the first 500 readings, because when looking at the temperature readings over that time period the two trees had vastly different

number, which did not occur at any point later in the data (see Figure 1). Additionally, during this time the incident PAR sensors were producing a lot of numbers well outside of the normal range.

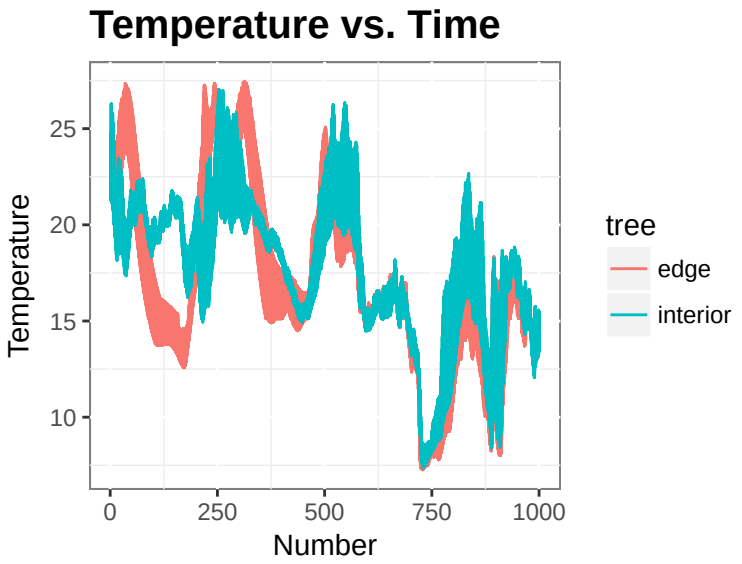

Figure 1: Difference between the two trees for the beginning of the data

Throughout this process the first thing I discovered were many gross outliers in temperature (below 0 and above 35), which were well outside the range of temperature during a summer month in California. Then looking at temperature and time I found that nodes 2, 3, 59, 64, 78, 123, 141, and 145 began to produce erroneous readings before dying, so I deleted all observations for those nodes once the issues arose. I also found that nodes 118, 122, and 196 began measuring humidity poorly at different periods along the way and choose to remove the observations after that time. There were many humidity observations over 100%. I choose to keep the measurements that were above 100, but remained close because they tracked with the other sensors, and made sense with the corresponding temperature and incident PAR at those times. However, I removed the observations above 110%, because there was no correlation with the other observations of humidity at that time. Nodes 135 and 40 failed to produce any useful incident PAR readings so they were also removed. Finally, I found an outlier in temperature and humidity that had humidity above 70% when the temperature was 22.2 and no other nodes had humidity above 40%, so I choose to remove that observation.

Throughout this process I incorporated my domain knowledge to understand to ensure that nodes at lower height had lower temperature, PAR values, and humidity. After this processing and reviewing the plots to make sure the variables made sense over time and space I determined that I had removed the gross outliers that did not reasonable correlate with the remainder of the data or my domain knowledge.

After the redwood source data was cleaned I combined the data set with the mote location data, which contained information about the height and direction of the nodes, and the dates data. The dates data was important, because the data stored at the node log did not record the date and time.

2.3 Data Exploration

I found that much of what I did to determine outliers brought up numerous questions that I attempted to explore by searching in smaller time units and finding ways to view the data in multiple dimensions, in addition to re-examining the scatter plots and I produced in order to clean the data.

After plotting the variables over time and adding transparency I noticed that there are less and less data points as time increases. Clearly this occurs because of the nodes and observations that were deleted over time in addition to nodes that stop working throughout the study, but I wanted to know specifically how many nodes remained towards the end of the study. After coloring by node ID I noticed that by the last ten days there were only a few nodes remaining (see Figure 2). This encouraged me to look into the data closer to the beginning of the study. At those times there are more nodes across height

and direction to allow an investigation of how those variables effect temperature, humidity, and incident PAR.

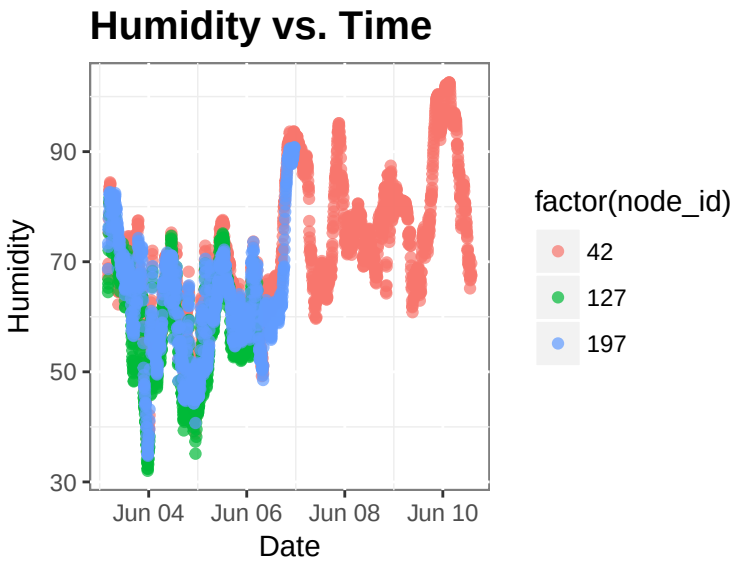

Figure 2: Declining number of nodes

Next, I choose to look into how the time of day effected temperature, humidity, and PAR values in order to ensure that temperatures and PAR were peaking in the middle of the day and that humidity was decreasing. To do so I plotted the variables over time and colored by hour of the day. After zooming in on a windows of three to four days I noticed that there were times when it look longer for temperature to reach peak value. I then looked at incident PAR, reflected PAR, and humidity and validated that PAR values also peaked later and humidity began to decline later. I also checked looked further into the expected inverse relationship between temperature and humidity. Additionally there were days with many low incident and reflected PAR readings, so I checked that corresponded with low temperature and high humidity, which seemed to be the case. Throughout this process I noticed that as temperature increases the variance of between nodes gets larger (see Figure 3)

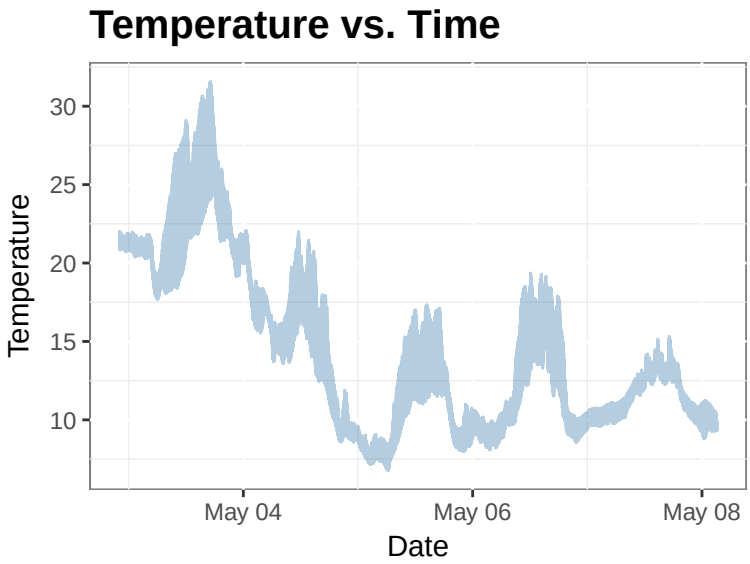

Figure 3: Increasing in variance with increasing temperature

In looking at how height effected the variables, both by coloring the time series with height and plotting height vs. each variable with node ID colored I ensured that values at the different heights were acting

as expected, but I was curious if direction was confounding, so I looked at how direction and height were related. First I investigated a histogram of height colored by direction and then scatter plots that of variables over time colored by direction and the size of the points representing the height. I then saw that it seemed that many of the nodes that were at the top of the trees and hence had the highest temperature and PAR values were primarily facing southwest. I attempted to make many plots to look into this phenomenon (including Figure 4) that incorporate height, direction, and incident PAR to show this, but was not able to present the information in a concise manner.

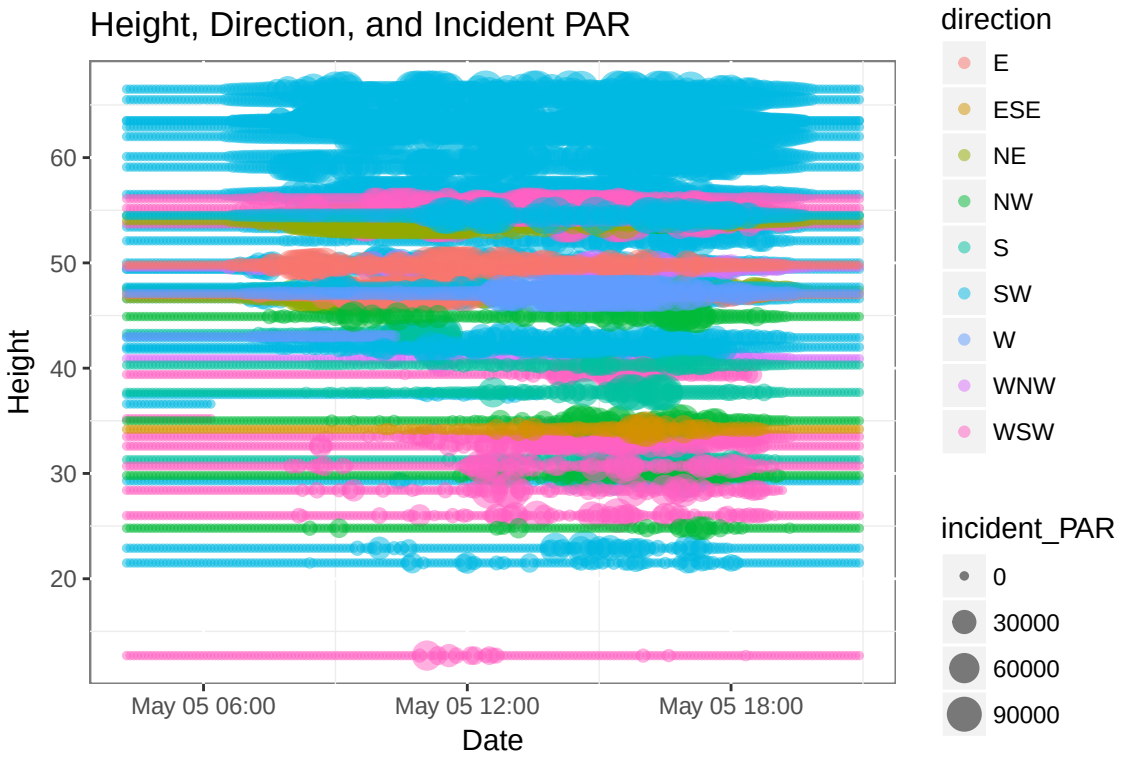

Figure 4: Attempting to display height, direction, incident PAR, and time on the same plot

In thinking further about direction, I wanted to looking into how the sun rising in the east and setting in the west effected incident PAR and reflected PAR. I was hoping to find that eastern facing nodes had higher PAR readings earlier in the day. I found the height of the node was a much more powerful indicator of the PAR values than direction. However, in looking into temperature I noticed that eastern facing nodes peaked in temperature sooner than the other nodes regardless of height. I also saw that there appeared to be three separate peaks of temperature each day. By looking into direction it seemed that eastern nodes peaked first, western and northern nodes second, and finally the southern facing nodes. These facts make sense with the path of the sun throughout the day.

3 Graphical Critique

In the Tolle et. al paper they present a number of graphics. Figure 3 is a set og plots including histograms for each of the values (temperature, humidity, incident PAR, reflected PAR), and box plots for each value with time, each value with height, and the difference in means of the values. These plots are useful in giving the readers a basic idea of the distributions of the values and the relationship of the values with height and time, providing a comprehensive view of the data. However, by placing all sixteen plots on a single page it is difficult to see the nuances of the relationships. It forces the font along the axes and titles to be too small to read easily. Additionally, the color choices with dark blue of the histogram and the much brighter box plots is off putting. Also, the box plots with reflected PAR and time is difficult to read considering that most of the values are concentrated around 0. It may have been better to choose a different method of presenting this information, such as a scatter plot, to see the patterns more distinctly.

Figure 4 plots the values over the course of a day and looks at a snapshot of each value in the middle of the day. In the snap shot the points are colored by direction. This figure is designed to incorporate many dimensions into a single page. Temperature and humidity plots incorporate the value, time, node ID, direction, and height, while incident and reflected do all of those except node ID. This is a concise way to look at a lot of information at one time, and there are many stories that can be told from looking at this figure as is reflected in the paper. Although, there are a few issues with plot. First, the coloring with respect to node ID in the top plots seems unnecessary. We do not get any additional information from that coloring, because there are simply too many nodes to identify a particular color and track that throughout the day. The color choices in the figure are also disconcerting, the bright green choice for the incident PAR and reflected PAR is too abrasive and clashes with the blue and pink used in the plots on the right. Using the arrows to represent direction was a creative idea, however it is difficult to see in these small plots and redundant as color is used as well. Finally the right plot of reflected PAR is too concentrated on the y-axis making it difficult to read what it happening.

4 Findings

4.1 First finding

The first finding I came to was that the variance of temperature and incident PAR increases as those values increase, while the variance of humidity increases as humidity decreases. I choose to only display this relationship for humidity, because I was able to fit the humidity values and variance on a single plot with out perturbing the scale of humidity as I would've had to do for temperature and incident PAR. This plot shows the variance of humidity at each sampling time and the actual value of humidity at that point. We can see that there is an inverse relationship between the two, but that there is a good correspondence between peaks in humidity with dips in variance and vice versa.

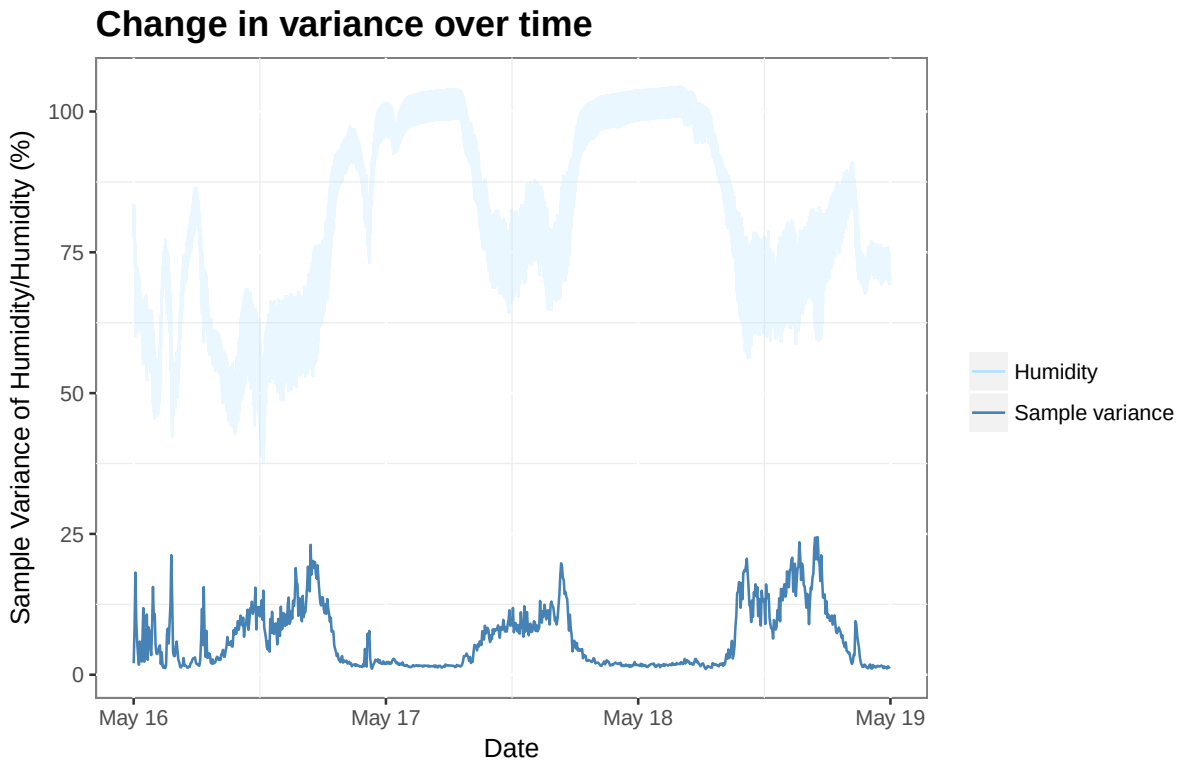

Figure 5: Plot showing how the variance of humidity changes as humidity changes

4.2 Second finding

The second finding is that there appear to be three separate peaks in temperature each day. This plot shows that those three peaks are related to the direction that the node faces. The nodes that face east, with the exception of node that faces east southeast peak first, followed by the northern and western nodes, including southwest and north southwest. While the southern facing nodes peak last. I choose not to display height in the plot because through my investigating I determined it effected the height of the peaks, but not the timing.

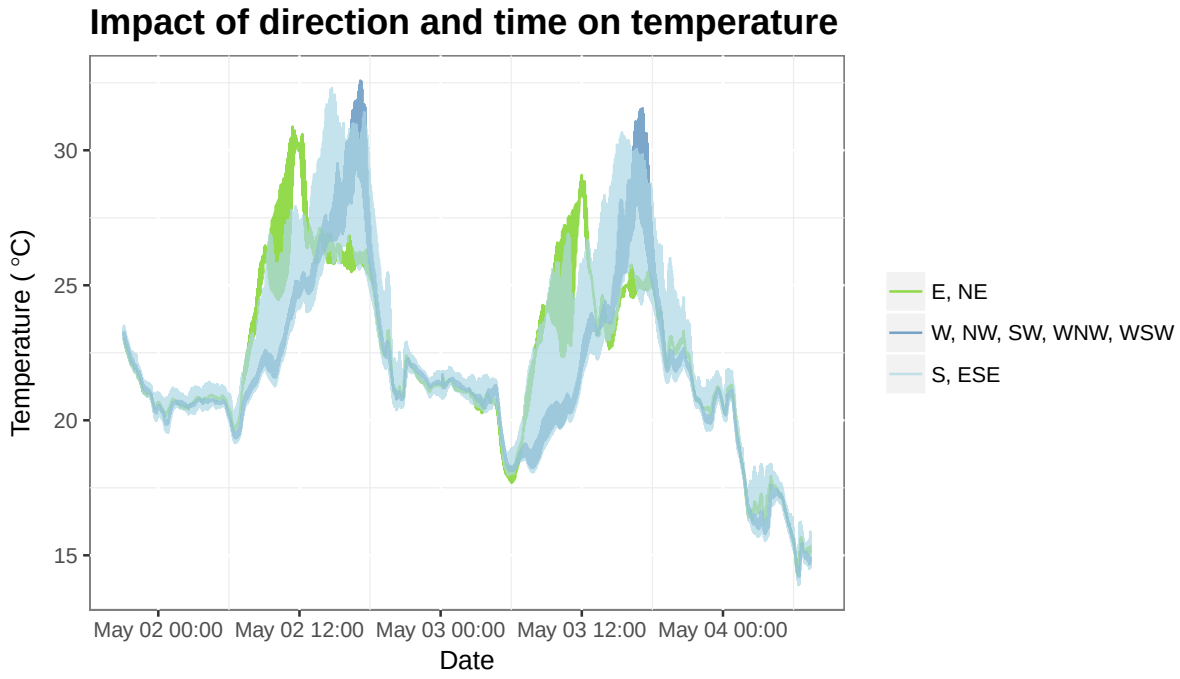

Figure 6: Plot that shows how the direction of the node effects timing of temperature peaks

4.3 Third finding

Finally, we look at a four day period where the middle two days where it appears the sun did not break through the fog until later in the day than is typical in this data set. We can see in incident PAR, which normally has a more rounded shape has a sharper edge on the left hand side as the sun comes up. The temperature on these days is lower and by coloring by hour of the day we can see that it rises slightly later in the day. Additionally the humidity remains higher later into the day than on a more normal day.

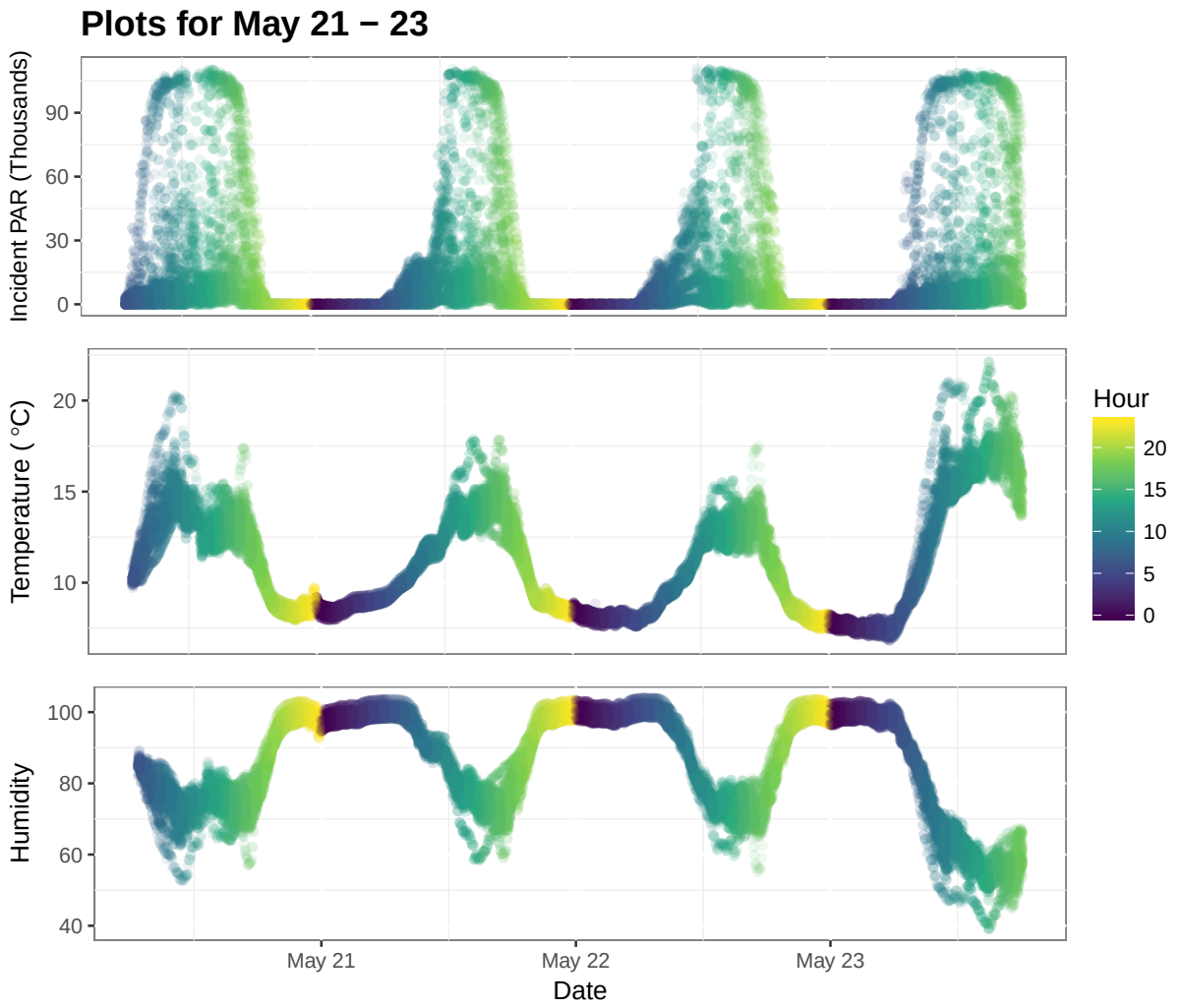

Figure 7: Plot investigating two days were the sun appears to come out later than other days by looking at incident PAR, temperature and humidity

5 Discussion

The size of the data was somewhat restrictive in both being too large and too small. If the data had been larger and there was more variance in the directions of the nodes and more trees there could have been more interesting findings hidden in the data. However, the sheer numner of data points meant that when producing plots to look at all the observations at once I had to wait for a period of time. This only created problems when I needed to update my plots continuously to look at different portions of data. The other time I ran into an issue with there being too many data point was when I attempted to creat an interactive plot with the Plotly package. I used too large of a subset of the data and it caused R to crash. Additionally due to the vast number of measurements per time period there was the potential for scatter plots to get too cluttered to be able to pick out what is actually going on. That meant that it was important to look at different subsets of the data to get a better idea of what was happening.

6 Conclusion

The data presented in "A Macroscope in the Redwoods." Tolle et. all presents an opportunity to discover how climatic variables relate to each, as well as how they vary with respect to space and time. The data collection method that lead to a large number of missing values and a number of erroneous measurements

made intensive cleaning of the data a necessity. Once that cleaning was done there was a good amount of information stored in the data, which painted a clear picture of a day in the life of a redwood tree.

References

[1] use this or generate it automatically with BibTex