Lab 1 - Redwood Data, Stat 215A, Fall 2018

9/13/2018

1 Introduction

This paper presents temporal and spatial analyses on temperature, humidity and photosynthetically active solar radiation (PAR) for a redwood tree using the data collected from a previous study from University of California, Berkeley [1]. Data were collected through a swarm of wireless sensor nodes scattered around the tree that logged data for more than a month at 5-minute intervals. Through this data, variations on the microclimate on this tree can be identified enabling possible research on complex environmental dynamics.

2 The Data

To enable dense data collection, a network of wireless sensors were deployed on a 70-meter tall redwood tree, both exterior and interior, starting at a height of 15 meters with a radial distance of about 0.1 to 1 meter from the trunk. Each battery-powered sensor node has 4 sensors to measure variables relevant to the study: temperature, humidity, ambient and direct PAR. There were design considerations for the sensor node enclosure to ensure proper protection against the elements. The sensors collected data once every 5 minutes for 44 days during the early summer, April 27 to June 10, since it was identified that most dynamic microclimatic variations happen during this period.

2.1 Data Collection

Each of the sensors in the node were calibrated before deployment. For the temperature and humidity sensors, a controllable weather chamber was used so that the sensor values can be calibrated to the correct, or at least near to the correct, set of values. The authors of the paper also highlighted that there still might be issues after calibration which will be attributed to the actual sensors themselves. Possibly, wrong data can be saved due to some electrical factors, such as interference, as seen in the data having some maxed out values and weird outliers not consistent with the trend. On the other hand, the PAR sensors were calibrated on top of a roof in direct sunlight for 2 days. This allowed variations in sunlight amount and angle be recorded to check the sensor functionality with respect to a high-quality sensor reference. My only issue with these calibration steps was that it was not calibrated in the field as a double check. When the sensor nodes are deployed on the tree, there are other factors that can affect the reliability of the sensor node data, such as shading from the foliage affecting sunlight sensing and wind affecting temperature and humidity sensing. This can explain some of the outliers in the data that will be presented in the next section.

Every 5-minute data was collected in two ways: through wireless communication to some server, and through offline storage on a 512kB memory per node. This redundancy is especially important since the wireless communication might fail sometimes either due to wakeup time synchronization or signal issues. In the end, there were more than 400,000 total samples collected from all sensor nodes, and only a quarter of which coming from the data transmitted wirelessly. I also have to note that while there is redundancy, mismatch between sensor node data can still appear. That is, the logged sensor data from a specific node at a specific time epoch might not be the same as the transmitted one. This might be due to two possible reasons: communication error due to contention on the wireless channel that the nodes are using, and hardware/electrical issues (especially due to the environment that the sensors are operating at) that can destroy some of the data being saved in local memory.

Due to different possible causes of data inconsistency, the data must be manually checked for errors, usually in the form of outliers. This process is described in the next section.

2.2 Data Cleaning

The collected data, both from the network and the logs, is raw and contains a lot of measurement errors. While it is not a guarantee that all wrong data can actually be removed (since some points might look like a true data based on a trend), we can minimize these by removing all obvious erroneous data. This will be primarily done on a time-based trend analysis. Each variable will be plotted against time and spot any data points that does not make sense given the context of the deployment environment, such as negative temperatures and humidity, data that doesn't change through time, and data that drifts too much after some time window. The steps for the data cleaning method that I did are as follows.

2.2.1 Outlier Rejection

The original authors presented their own outlier rejection scheme in their paper. They removed sensor readings that are outside the normal operating range and those that are inconsistent with the other sensors. The typical problem that they have identified is that sensor data tends to be erroneous as the battery level of the node decreases. However, their approach was to remove all node data from the dataset having voltage level outside of 2.4V-3V. This is quite problematic since there is a chance where the node voltage sensor is just faulty and not recording the correct values. If only the voltage sensor is wrong, then the data collected by that sensor node will be wasted if it is removed. Therefore, the outlier rejection scheme that I adapted was more conservative by removing only the datapoints on the variable/sensor data that is producing inconsistent results and not the whole sensor node itself. This will lead to more data being preserved which can be used for analysis. While it can be argued that their outlier rejection scheme might turn out to be correct, I will prioritize having more data to be used for analysis than removing samples completely.

I started my outlier rejection scheme by plotting each of the variables versus time. This is based on the assumption that there should be an evident trend on these variables and that this trend should be almost the same when different sensors are considered. I did the outlier removal on both the logged and network datasets separately.

For the logged dataset, plotting the voltage allowed me to identify nodes 128, 134, 135, 141, 142, 143, and 145 having a constant voltage all the time. However, when considering just these nodes, all other variables follows some trend also seen for the other nodes with correct voltage reports. This implies that only the voltage sensor is malfunctioning while the other sensors are alright. I simply removed the voltage values for these nodes to keep the other data present. Plotting the humidity shows negative values all belonging to node 29. Again, after checking the value of other sensors to be alright, I just removed the erroneous humidity points and kept the rest. The ambient PAR (hamatop) and direct PAR (hamabot) showed node 40 being an outlier so I promptly removed that value as well. After passing through all the variables and removing the outliers, I double checked the plots of the time series and made sure no other extreme outliers are present.

For the network datset, I also started by plotting the voltage time series. This shows node 134, 135, 141, 145 having constant voltages so I removed the corresponding voltage values and keeping the rest. Interestingly, these nodes are just a subset of the nodes identified to have erroneous voltage values on the logged dataset. The other nodes present earlier were not in the network dataset, meaning the nodes failed to transmit their values to the server. Plotting humidity shows node 78, 123, and 141 having negative values. However, the negative values were only present on the latter part of the epoch. This means that the identified nodes were functioning properly initially, until problems occurred (possibly due to low battery, as identified in the paper). Instead of removing all of the data associated with the nodes, I only removed those occurring after some epoch: 3190, 4917, 9001 for nodes 78, 123, and 141 respectively. This identification of epoch relied too much on visual judgement on when to call a specific time to be the time corresponding to the start of errors. Nodes 145 and 3 showed outlier behavior while plotting the temperature after epoch 3500 and 3775 respectively. Both node data after those specified time points were removed accordingly.

Again, it is good to reiterate that the process of removing outliers were done visually through the aid of time series plots. There are limitations with this approach. I was not able to check the plots of different variable pairs to see if there are outliers. While this can be done, it would take a much longer time to inspect each of

the plot combinations. There is a concern that I might end up cleaning the data too much to the point that no interesting relationships can be observed in the end. Therefore, limiting outlier detection to time series plots are deemed as a good trade-off. Another limitation of this approach is the fact that the data needs to be inspected. If there are significantly more variables present in this dataset, the amount of time performing data cleaning would also increase. There is really no other possible option since I have no prior knowledge on what the data looks like and thus cannot determine outliers automatically. At this point, since the outlier detection was done through inspection, the process becomes more subjective than objective, and this is the true limitation of this procedure in general.

2.2.2 Dataset Merging

After sifting through all the data from both sources and removing points considered as outliers, the next step was to combine the two datasets together. There was initially an option to either consider both datasets separately or not. However, since these datasets are actually coming from the same set of nodes, there is really no reason to separate them. Redundant data that would appear because of the merge should not significantly affect the results. Moreover, these redundant data can also be easily removed by just checking matching node IDs and epochs.

One interesting thing that I did for the merge was to make the voltage values on the same scale. The network dataset contains values that are in the 200-300 range, while the logged dataset records the actual voltage values (around 2.4V to 3V). This means that the server collecting the data from the wireless transmission is not doing any conversion from the raw voltage sensor data to the true voltage value. To equalize the voltage scales, I first created a scatter plot of the voltage levels from the nodes that are present in both the datasets. With this, it became evident that there is almost a linear relationship between the two voltage values. I did a linear fit on this plot to obtain the coefficients that will then be used for scaling the raw voltage values.

Another thing that is interesting to note during dataset merging was the presence of different values for variables having the same node ID and epoch. This does not make sense since the values are coming from the same node at the same time. This anomaly can probably be explained by transmission errors through the wireless network. Luckily, there were only a few cases of this happening and taking the average between the two different values is good enough and does not mess up the trend when the time series is plotted.

2.3 Data Exploration

Now that the dataset has been cleaned and merged, we can now ask for questions about the data. Perhaps the most obvious question would be: how does humidity, temperature, ambient and direct PAR varies through time at different heights? Can we draw conclusions on how these variables vary both temporally and spatially?

The following figures in the next page showcases the variation of these parameters both in temporal and spatial domain. There are few observations that can be made here. First, there's a large concentration of nodes near the bottom of the tree contributing data for the first quarter of the total epoch (around 10 days). However, as time goes on, fewer and fewer nodes are contributing data, as can be seen by the gradual effect of transparency being evident. Additionally, it is the top nodes are remain operating until the end of the data gathering phase. Another thing to note is the lack of overall periodicity of the humidity and temperature data. There are notable mini peaking in the data which actually corresponds to the days and nights. However, looking at the whole trend, it does show some dynamic unpredictable trend, showing the presence of some micro-climate. In contrast, there is significant periodicity in the ambient and direct PAR. The peaks basically just tracks the mornings when the sun is up, which makes sense. Moreoever, it can also be seen that the bottom nodes does not contribute much to the peaks of the PAR data, which also makes sense as these bottom nodes are more likely to be covered by foliage and does not receive sunlight that much.

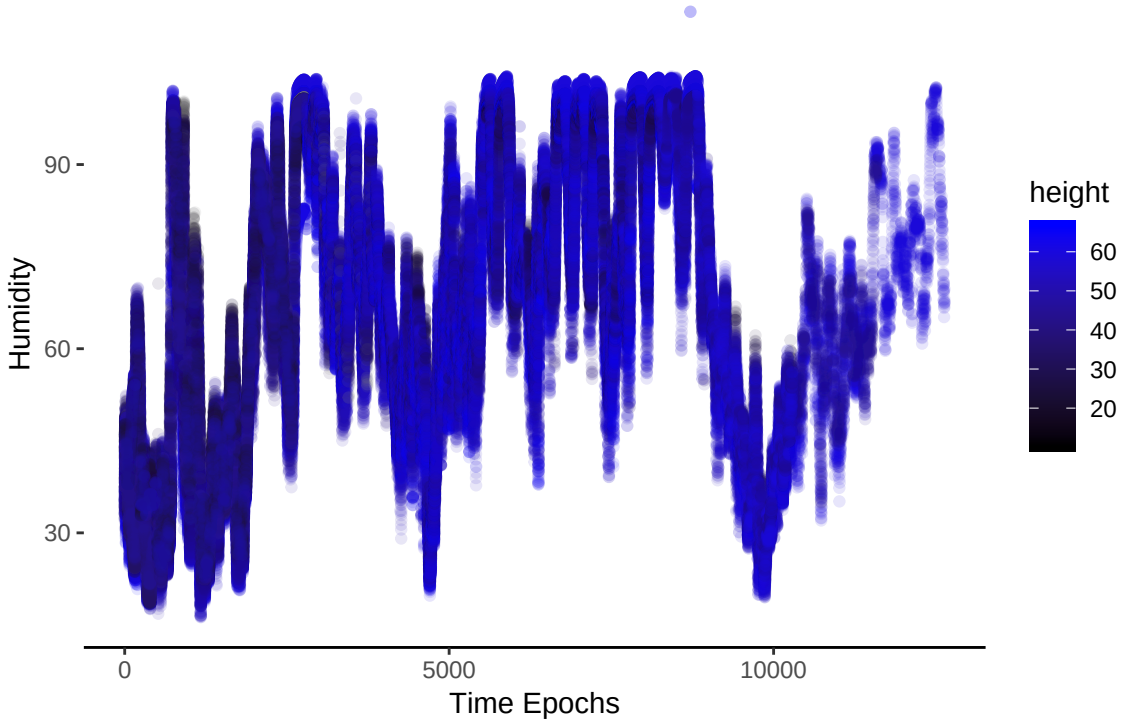

Figure 1: Humidity variation across time and node height

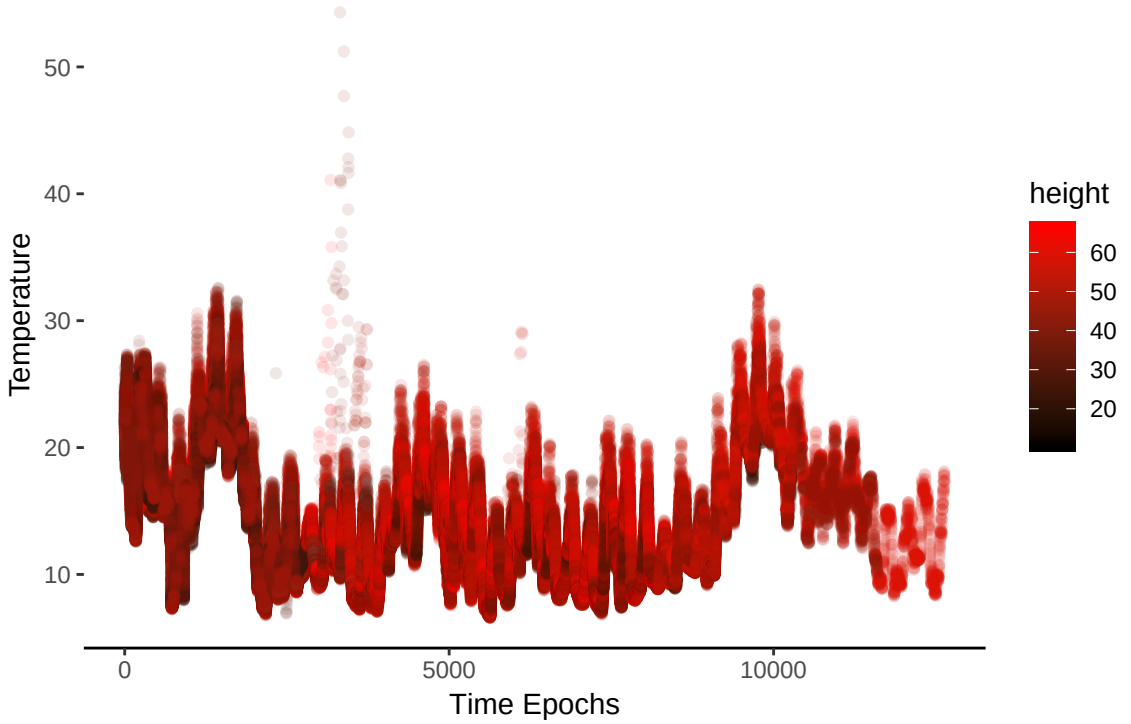

Figure 2: Temperature variation across time and node height

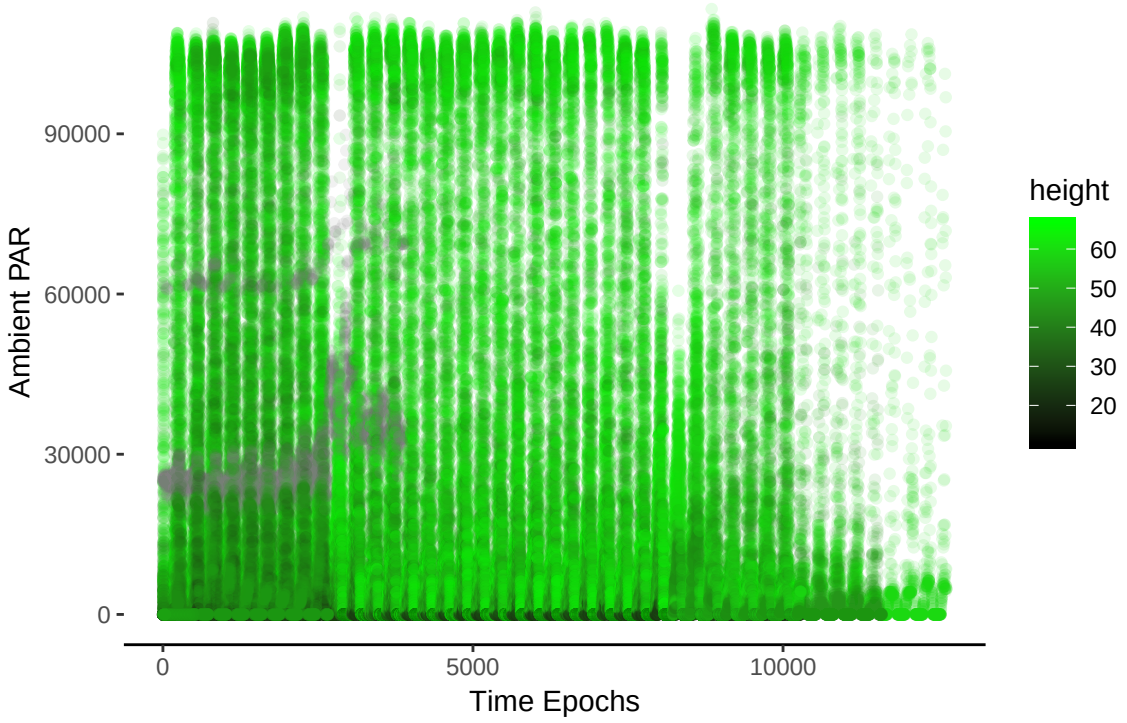

Figure 3: Ambient PAR (hamatop) variation across time and node height

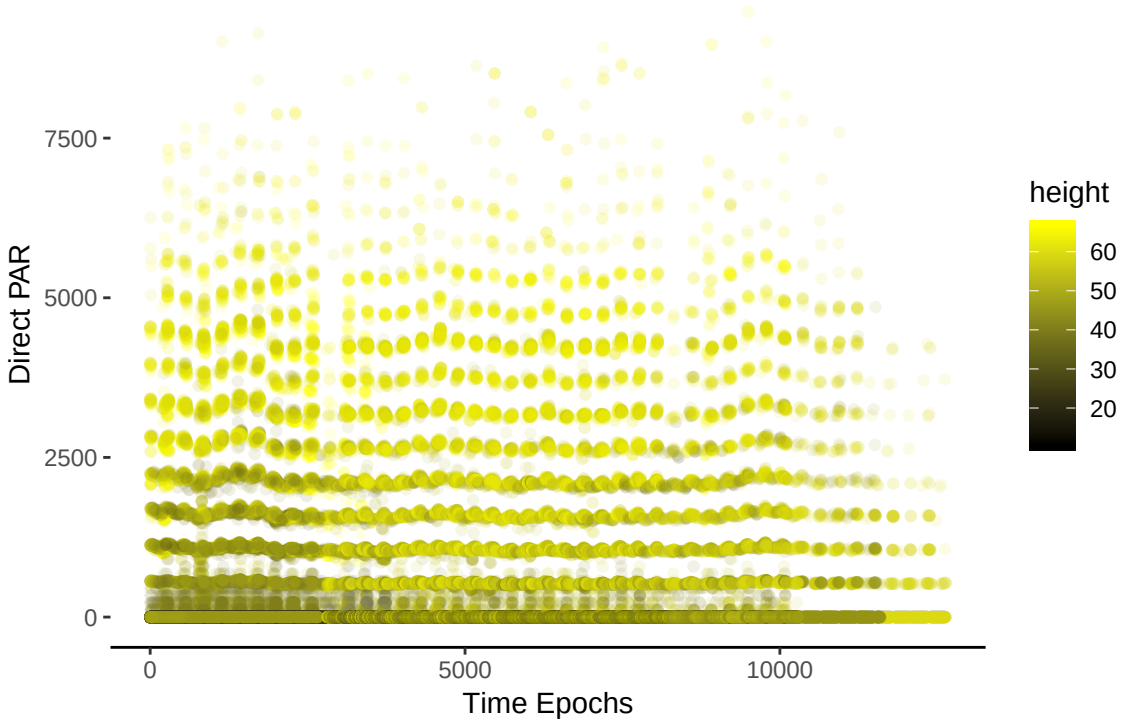

Figure 4: Direct PAR (hamabot) variation across time and node height

3 Graphical Critique

Looking at the original paper, they have provided 2 main figures for their analysis. Figure 3 presents the distribution of the data plotted against time and space, while Figure 4 presents the actual variation of the data on a 1-day period as well as the node height. We will criticize these figures for their ease of interpretability and readability.

Figure 3 tries to present a lot of information using a lot of sub-figures. Interpreting the series of plots is easy due to the presence of axis labels. However, the choice of plotting method (especially for b to d), leads to a lot of lines being presented in a small sub-plot, making it somewhat unreadable. Moreover, the y-axis for the node heights are too compressed and are not even in the proper scale. This could have been fixed by creating separate larger plots such that the authors can additionally add information as they wish. Alternatively, Figure 3 can aslo benefit a lot from 3D plotting to combine information regarding the sensor node data in both space and time. Some of the plots can actually be merged. For example, 3b and 3c can be combined to highlight the variations of temperature, humidity, and PAR, for different node heights (on different colors) across the whole time.

Figure 4 is a better plot showing the information that the paper wants to focus on. Plotting all the sensor node data across time is a good idea for me. However, since the height information is on a separate plot, then it is hard to give a conclusion on how node height affects the temperature and humidity through time. Moreover, looking at the figure alone, the node height scatter plot also highlights 3 nodes but does not explain the importance of those nodes. Figure 4 also includes incident and ambient PAR measurements but did it in a scatter plot instead with a line fit. The problem of having a separate plot for time and space is also apparent here: it is hard to identify whether node location matters in the variability of PAR through time. Moreover, color choice could have been better: the use of bright green is somewhat hard for the eyes and might not print well on grayscale. Scaling for the height-based scatter plot is also problematic since much of the points stick to the left-hand side of the plot. Lastly, looking at the plots alone, I am not sure why a specific timestep is being highlighted as it does not seem to be an outlier or anything significant happening at that point.

All in all, the plots would benefit from addition of more colors to introduce more information on the plot without sacrificing readability. Bigger plots with properly-scaled axes would also be a good thing to incorporate as well.

4 Findings

4.1 First finding

The data exploration stage gave a little bit of an idea that the number of sensor nodes contributing data diminishes as time goes on. This can be attributed either due to environmental effects or decreasing battery charge during the deployment time. The following plot shows a histogram of the total number of nodes that are contributing data (both logged and transmitted). Aside from confirming our observations earlier, it can also be seen that there are two epochs that triggered a massive decline in the number of sensor nodes that's holding data. Specifically, there was something that happened on time epoch 2500 that rendered half of the total sensor nodes inoperable, and another at near time epoch 10000. Looking back to our data exploration plots, there were nothing significant being measured near those points of interest. The most plausible explanation would have been some environmental effect such as a strong rain which can make water penetrate the sensor node enclosure. Between these two epochs, however, the steady decline of sensor node numbers can be seen, which is most likely due to decreasing battery voltage.

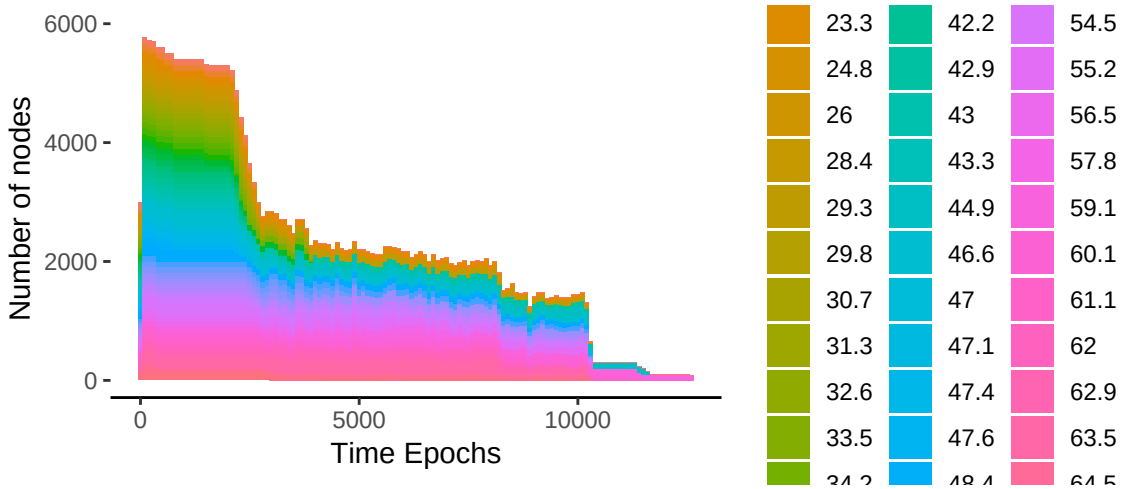

Figure 5: Histogram of node activities at different deployment heights through time

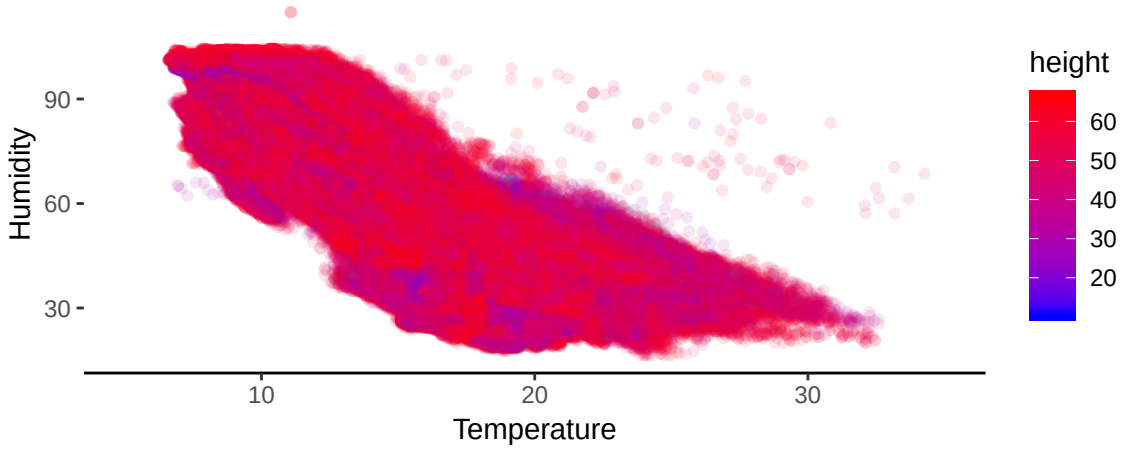

Figure 6: Spread of Humidity versus Temperature for different node heights

4.2 Second finding

Another finding that I am interested in is how correlated the temperature is to humidity. It can be seen from Figure 6 that there is some obvious negative correlation. However, there is some significant spread on this relationship. Moreover, this large spread is exhibited by all nodes from different heights. An interesting graphic that can be included here is an interactive plot showing how this humidity versus temperature spread varies per epoch. That would have been useful to identify the presence of micro-climates within the tree. I have yet to explore this functionality and would be great as future work.

4.3 Third finding

The third finding that I want to explore is the correlation between the ambient and direct PAR across all nodes. The resulting plot was particularly interesting. Going back to the exploratory plots, the only information that can be acquired from those plots where the correlation of the PAR peaks to mornings where the sun shines. In this finding, however, we see direct PAR values that are in discrete steps across all values of ambient PAR. This might imply that the direct PAR sensor is erroneous and is giving off some default values.

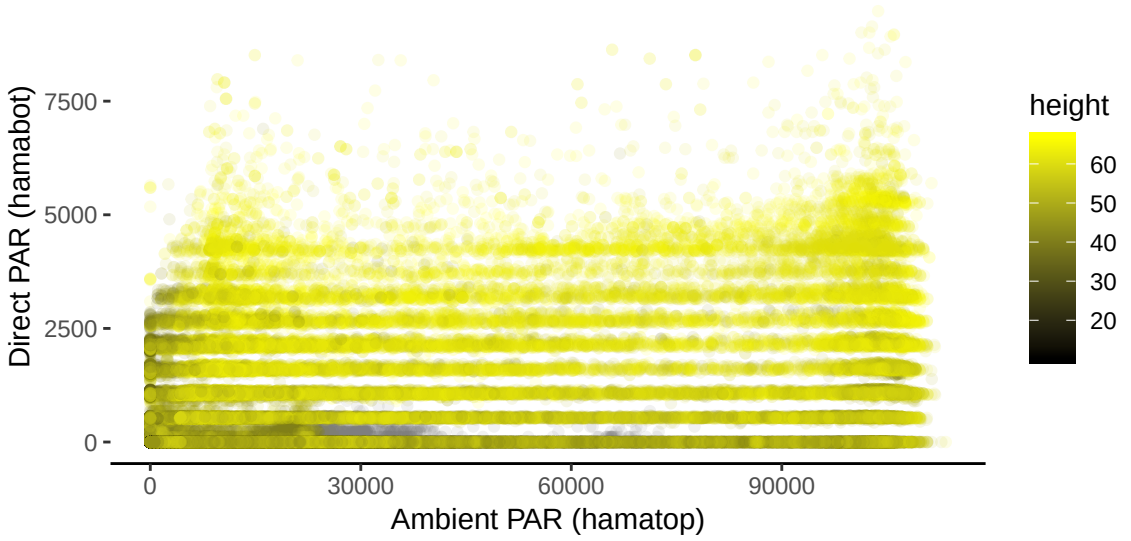

Figure 7: Correlation between Ambient and Direct PAR

There is a need for a more understanding on how these sensors actually transmit data to the microcontroller.

5 Discussion

This is an interesting dataset due to the size and the fact that there are two, albeit almost redundant, sources of data. However, handling data this large is not really new to me. Given the limited number of variables, data exploration and visualization can actually be performed to test and see different relationships on different variable combinations. For my exploration, I have focused on the data points plotted against time and did my data cleaning through that. For the research that I am doing, I am used to identifying relationships and outliers versus time, and not with the different combinations of exploratory variables. Maybe, next time I should focus on this as well as I might have missed a lot of important relationships that I could have included in this report.

6 Conclusion

This paper basically presented some preliminary findings on the redwood tree data. Even though there is a large amount of data, there were some outliers which had to be conservatively removed by visual inspection by plotting the time series per variable. The advantage of that scheme is easily identifying some trends that occur on each sensor value through time. However, it is very tedious and not scalable for projects having significantly more data. Moreover, it is highly subjective and depends on the one looking at the plot to determine whether something is indeed an outlier or not. Several exploratory plots and findings were then presented after post-processing. Given the size of the dataset, there are a lot more possible analysis that can be done in the future to extract information regarding the redwood tree's microclimate.