Project 1 - Redwood Data Report

Johnny Antoun (0679537) Jose Pliego (2716768)

September 23, 2021

Contents

1 Data Collection

1.1 Paper Summary

The paper by Tolle et al. (2005) presents a case study in which a wireless sensor network is used to record 44 days in the life of a 70-meter tall redwood tree. The redwood tree is selected as an interesting species to study as it is known to have substantial variation and to have significant temporal dynamics. The case study is unique as it involves gathering a data set that was previously not obtainable by making use of new technology, a wireless sensor network ("macroscope"). Previous set-ups consisting of limited apparatus had confirmed that there was variation across the tree but failed to capture a detailed picture of the entire structure over time. By the end of the moth, an analysis of system performance data was performed in order to make future deployments' results more accurate. Several lessons were learned through the initial deployment.

First, when the sensors get small enough and the phenomenon gets directional enough, tiny differences in positioning get magnified into large effects on the resulting data (especially noticeable in PAR data). During a clear day, each mote's readings fluctuated leading to the belief that foliage was blocking solar access to motes but the patterns were consistent between different days. Slightly different orientations for each light sensor resulted in different fluctuation patterns for each node, yielding the seemingly "random" appearance of the light data. The noisy data was actually a deterministic response by a highly focused sensor.

Second, the success of a deployment depends crucially on the management of the network. Any long-term sensor network deployment should include a network monitoring component that can provide real-time information about the performance of the system, and can alert the researchers when the system begins to behave abnormally. The network can then provide a means to detect and compensate for failures in the logging, while the logging provides a means to compensate for failures in the network.

Third, having verified the existence of spatial gradients in the microclimate around a redwood tree through the deployment, and captured data over time, the data can then be used to validate biological theories. For example, plant biologists can build a quantitative model of the effects of microclimatic gradients on the sap flow rate using obtained data to quantify previous knowledge of the impacts of humidity, air temperature, and PAR on sap flow rate.

1.2 Data Collection Process

Gathering data on the environmental dynamics around the redwood tree involved careful system design and deployment methods. Before placing into the field, two calibration checks (roof and chamber) were performed. Roof and chamber calibration checks confirmed robust performance for different subsets of the used sensors. The roof calibration allowed the researchers to establish that PAR sensor readings were acceptable. The chamber calibration involved a two-point calibration to obtain accurate results for humidity and temperature.

Once in the field, the electronics used needed to be protected from the weather while safely exposing the sensors. In terms of time frame, information is gathered for sensors every 5 minutes during a month in the early summer, which contains the most dynamic microclimatic variation. The nodes were placed on the west side of the tree (thicker canopy provides protection from environmental effects) from 15m above ground level to around 70m, with roughly 2m between nodes and at a radial distance of 0.1m-1m from the trunk. In additional, several nodes where added outside of angular and radial envelope to measure microclimate in the immediate vicinity.

The choice of measured parameters was driven by the biological requirements. Traditional climate variables were measured like temperature, humidity, and light levels. Temperature and relative humidity feed relate to transpiration. Photosynthetically active radiation (PAR) provides information about energy available for photosynthesis and gives information about drivers for the carbon balance in the forest. Total Solar Radiation (TSR) was ignored as the sensor was too sensitive and PAR was being measured. Moreover, barometric pressure was excluded as it is simply too diffuse to show important differences.

To provide a backup in case of network failure and to establish a basis for analyzing the performance of the net-work, the researchers extended the TASK framework to include a local data logging system. The data logger recorded every reading taken by every query before the readings were passed to the multi-hop routing

layer, and stopped recording once the 512 kB flash chip was full. After the deployment, they attached each mote to a serial connection, and then installed a new program to transmit the contents of the flash over the serial link. They chose to include a complete data logger because they knew that the capacity of the flash was sufficient for the duration of the deployment.

2 Data Cleaning

By looking at the histograms for the different variables, we can see that there are some disparities in the files regarding the variable `voltage`. After reading the user manual (see *MPR/MIB User's Manual* (2004), p.23) for the MICA2DOT platform that was used in the study, we found out that the data retrieved from the network has the measurements from the ADC. According to the documentation, we can convert these measurements to battery voltage using the following equation:

$$V_{batt} = V_{ref} \times ADC\_FS/ADC\_Count,$$

where V_{batt} is the battery voltage, $V_{ref} = 0.6$ is the external voltage reference, $ADC\_FS = 1024$ is the resolution of the ADC, and ADC_Count is the value observed by the monitor. After doing this conversion, figure 1 shows that the readings are coherent between both data sets.

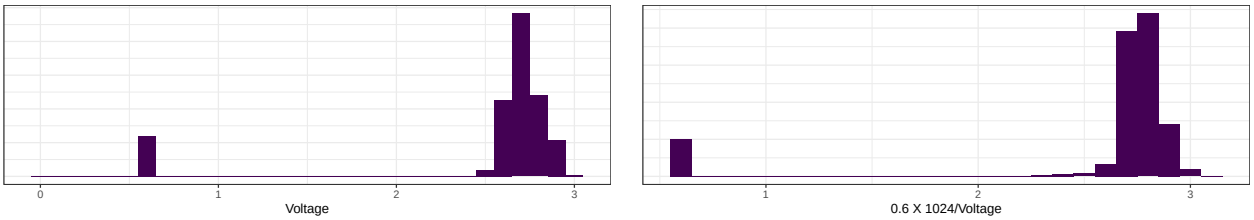

Figure 1: Voltage reading in the log (left) and network (right) data.

We talk about how we dealt with outliers later in this report. For now, we concatenate both data sets and filter out repeated rows, using the fact that `nodeid` and `epoch` together are a unique identifier for one measure. In the case when we have the same values for `nodeid` and `epoch` but differences in other variables, we try to keep the values observed in the data retrieved from the network. If that is not possible (because the repeated identifiers are in the log data), we average the values for the variables. No combination of `nodeid` and `epoch` appears more than two times, and only 94 combinations were repeated after discarding the variables that were not used in this project (`parent`, `depth`, and `humid_adj`). The resulting data frame has 319,031 rows and 7 variables.

Before dealing with missing values, we read the `sonoma-dates` file to pair each epoch with the date and time in which the readings were taken. This way we can identify missing values in specific nodes during certain periods of time.

The data set now has 2.77% of missing values in `humidity`, `humid_temp`, `hamabot`, and `hamatop`. Diving deeper into these missing values we see that most of them come from node 122, which failed to register data between '2004-05-07' and '2004-05-29.' Other missing values come from node 15 which failed to register data between '2004-04-30' and '2004-05-06,' and the rest come from node 128 which failed to register data between '2004-04-30' and '2004-05-05.' After removing missing values we lose less than 9,000 rows. The resulting data set has 310,179 rows and 8 columns.

After removing missing values, we read the location data for each node contained in the file `mote-location-data.txt`. We join this data to the full data set using the column `ID`, which we rename to `nodeid` so the names match.

We see that there is only one observation with `nodeid` 65535, which is clearly a mistake so we remove that observation. After joining with the location data, we have 310,178 rows and 12 columns (10 variables considering that `epoch` and `nodeid` are identifiers). Nodes 100 and 135 have no location data. We keep these

observations because they are useful for parts of the analysis that do not take into account the location of the nodes.

2.1 Outlier rejection

The data has plenty of outliers, many of which seem to be due to misreadings. One example is that there are many negative values for `humidity`. This variable is a percentage of relative humidity, meaning that it can be greater than 100% but never lower than 0%. After trying to identify outliers visually, we decided to follow the advice in the paper and filter out readings with voltage values lower than 2.4 or higher than 3. Tolle et al. (2005) mention that these boundaries show that the battery of a node is running out and thus the nodes yield unreliable data. Even though some of the readings under outlier voltages were not outliers in each variable, we chose to remove them so we do not risk performing the analysis with unreliable data. Figure 2 shows that by removing the voltage outliers, some of the humidity outliers are also removed (the value under -5000 was manually removed to make the histogram more readable). This phenomenon is mostly repeated across all the variables of interest.

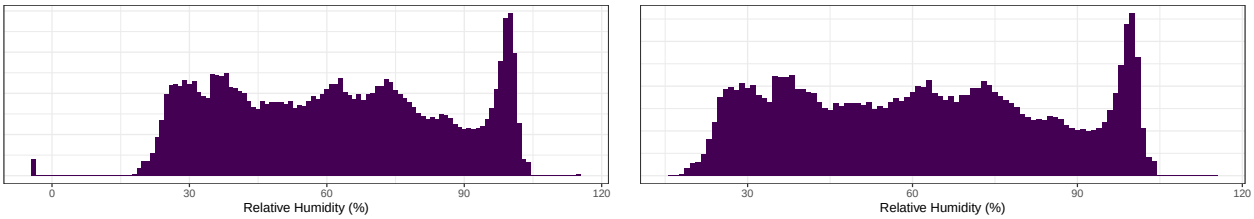

Figure 2: Histograms of humidity before (left) and after (right) filtering extreme voltage readings.

Now we can visualize quantiles and histograms to identify outliers in each variable. In figure 3 we see that the horizontal axis reaches up to 125. This is because of two readings that have values 117 and 122, while the rest of the temperature readings are under 33. It seems reasonable to remove these two rows because they are clearly mistakes or misreadings.

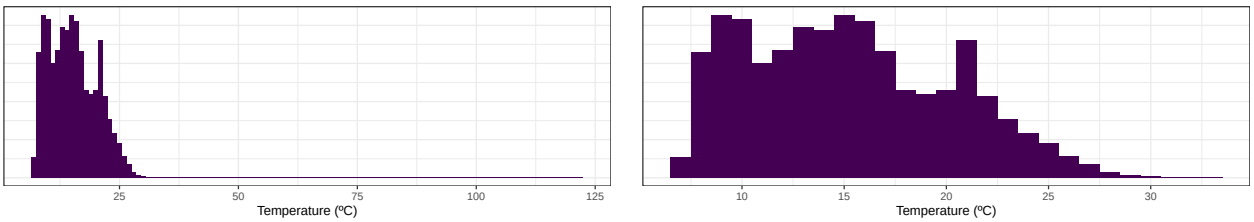

Figure 3: Histograms of temperature before (left) and after (right) removing outliers.

For the reflected and incident PAR, we have to convert the values to the units used by Tolle et al. (2005). In the data set we have Lux readings and we want μmol m^{-2} s^{-1}, so we have to divide the columns `hamatop` (incident PAR) and `hamabot` (reflected PAR) by 54.

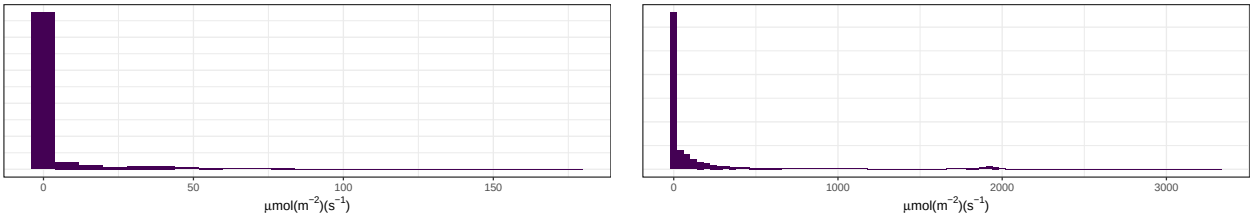

Figure 4: Histograms of reflected (left) and incident (right) PAR.

Figure 4 shows that the distributions of reflected and incident PAR are heavily skewed right. Reflected PAR has some values over 100 and incident PAR has some values over 2000. However, the histograms do not give enough information on whether these values are outliers or simply come from the heavy-tailed distributions.

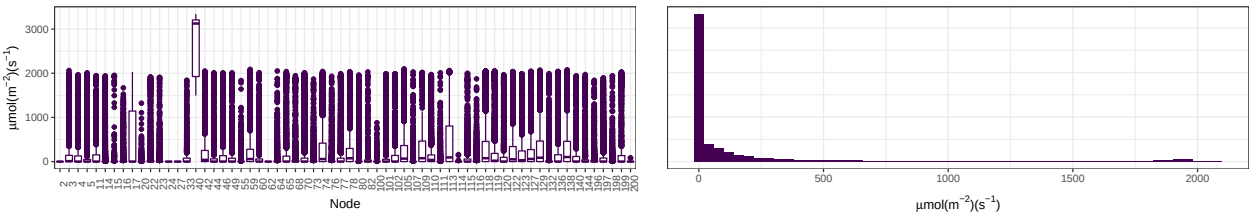

Figure 5: Distribution of incident PAR by node and after removing node 40.

Figure 5 shows that node 40 is giving extreme readings for incident PAR, not in line to what is returned by the rest of the nodes. We choose to delete the rows coming from this node because it probably was not well calibrated or malfunctioned during the deployment. The new histogram for incident PAR shows that there are no values over 2500.

Repeating this analysis by node for each variable does not show any obvious patterns, so we decide to work with the data remaining so far. The resulting data set has 277,058 rows and 12 columns with 0.8% of missing values in `tree`, `dist`, `direc`, and `height` corresponding to the nodes that have no location data.

3 Data Exploration

After looking at scatterplots of all the variables across different time windows, we decided to present two of them here. The time period chosen is the first week of data collection, which was chosen because some nodes began failing afterwards and because the correlation structure is similar across different weeks. Figure 6 shows that there is a positive correlation (~ 0.475) between incident and reflected PAR, which follows intuition because both variables are related to sunlight. However, due to the noisy reading for these variables, it is hard to tell if the true relationship is linear. Figure 6 also shows a strong negative correlation between humidity and temperature ($\sim$ -0.671). This is also expected since, as mentioned by Tolle et al. (2005), warm days are dry and colder days are more humid in California.

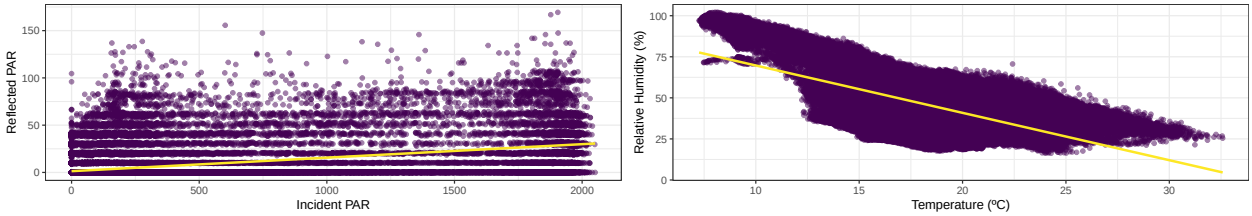

Figure 6: Pairwise scatterplots for incident vs reflected PAR, and humidity vs temperature.

To see if other variables are associated with incident PAR, we also plotted scatterplots against voltage, height of the node, and distance to the tree trunk. As shown in figure 7, There seems to be a linear relationship characterized by a positive correlation of ~ 0.305 between incident PAR and height. Intuitively, this relationship can be explained because higher nodes have more exposure to sunlight. The linear relationship is a bit more clear between these two variables, since low height nodes have almost no high values in incident PAR.

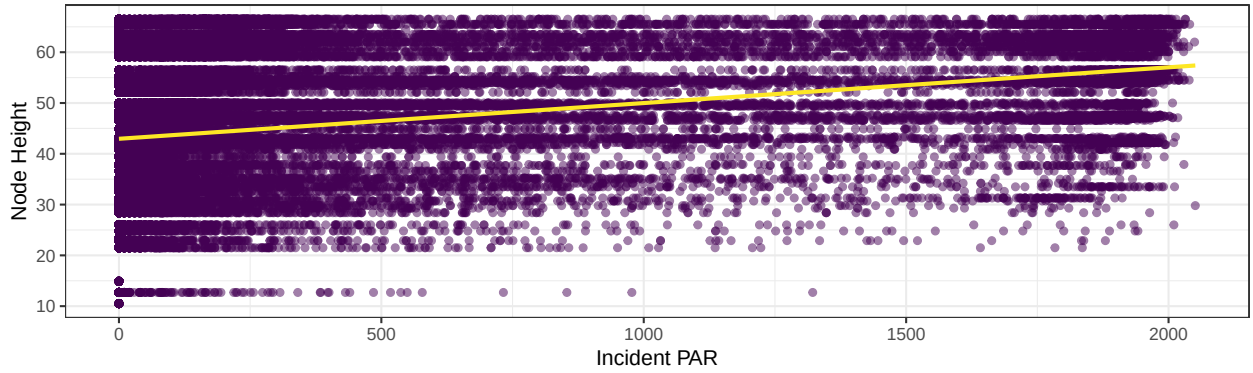

Figure 7: Scatterplot for Node Height vs Incident PAR.

To further illustrate our exploration, we include some time series plots using height as a color cue. Figure 8 shows the daily time series for temperature, humidity, reflected, and incident PAR. We decided to plot the daily mean to summarise the data. We also tried to use the median but the skewness of the PAR variables collapsed all observations to zero. In humidity and temperature, the structure was the same when using mean or median. In figure 8 we can see that higher node position is related to slightly higher temperatures, and reflected and incident PAR, while lower nodes tend to have slightly higher relative humidity.

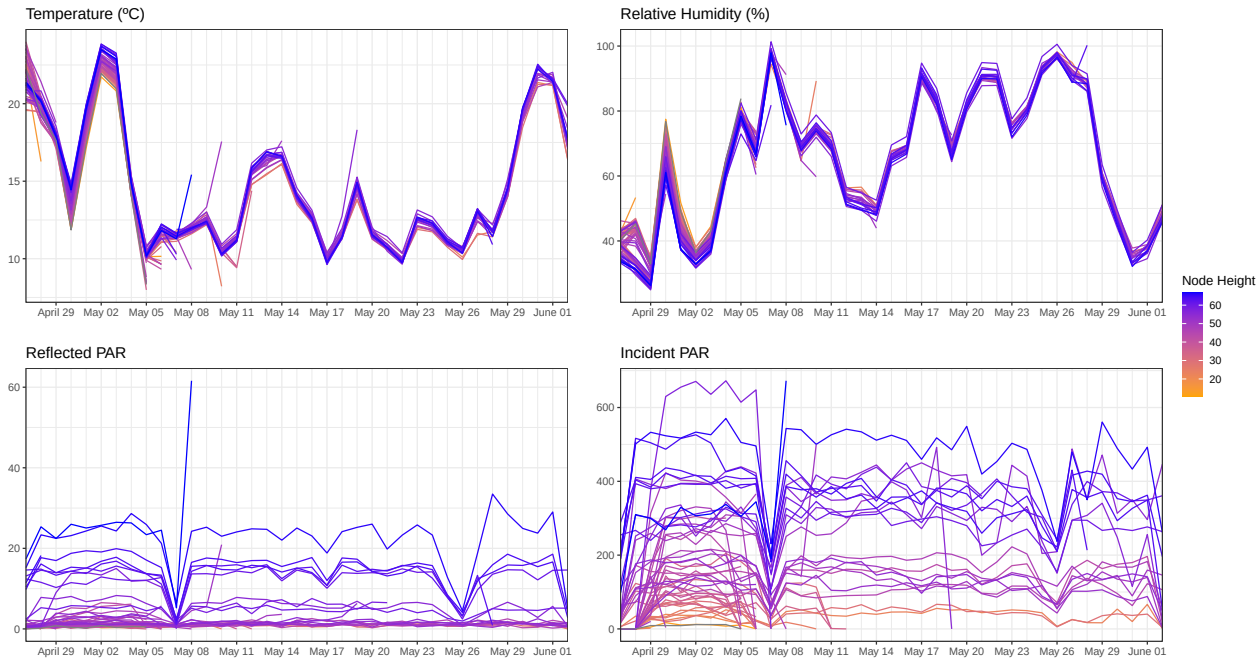

Figure 8: Daily time series for the study variables.

Figure 8 also shows some strange behaviors in the nodes. We see some truncated lines showing that we have no more readings for the corresponding nodes. We can also see that most nodes give an unusual reading before they are truncated. This unusual measurements were not detected during outlier rejection because they are evident when looking at the daily average by height, but not when looking at the complete distribution of the variables. We can also note two valleys in the PAR variables, corresponding to May 7th and May 26th. These two valleys most likely come from cloudy days in which the mean measured values shrinks close to zero.

To conclude this section we include a PCA of the data. For the PCA we include the four variables (humidity, temperature, reflected PAR, and incident PAR) along with the node heights and the distance from the tree trunk. We remove the rows for which we have no location information. We also center and scale the data before performing PCA. Figure **??** shows that more 75% of the variability in the data can be explained by the first three principal components, so this data set could be reasonably approximated by some low-dimensional representation.

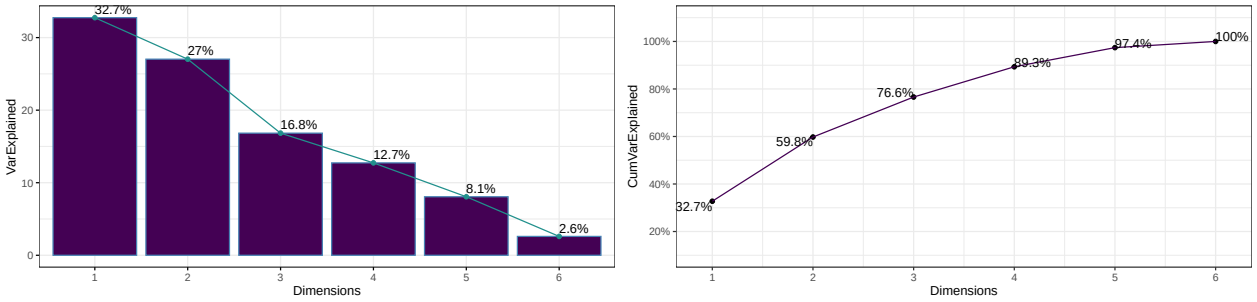

Figure 9: Screeplot and cumulative variance explained.

4 Interesting Findings

First, one of the main findings from the exploratory data analysis is that there seems to be a relationship between temperature and humidity, and height. We see in figure 8 that higher nodes tend to be associated with higher temperatures and lower humidity. This jumps out directly from the time series plots, while the original article does not see this relationship until they plot the differences from each timestep mean.

Second, in the time series plots we also found out that, even after removing outlier voltage readings, the nodes tend to give unusual readings just before running out of battery. This was not apparent in the original paper because these unusual readings are not unusual in the context of the overall distribution, but are outliers when looking at a given epoch for all the nodes.

Lastly, we performed a K-Means clustering analysis to confirm that humidity and temperature have some inherent relation to the height of the nodes. We compute the average temperature and humidity by height, apply the K-Means algorithm with two centers and then look at the distribution of height by cluster. If there is no inherent relationship between height and both humidity and temperature, we would expect the distributions by cluster to be completely overlapping. However, as seen in figure 10, the distributions by cluster show significant differences, with cluster 1 having the higher nodes.

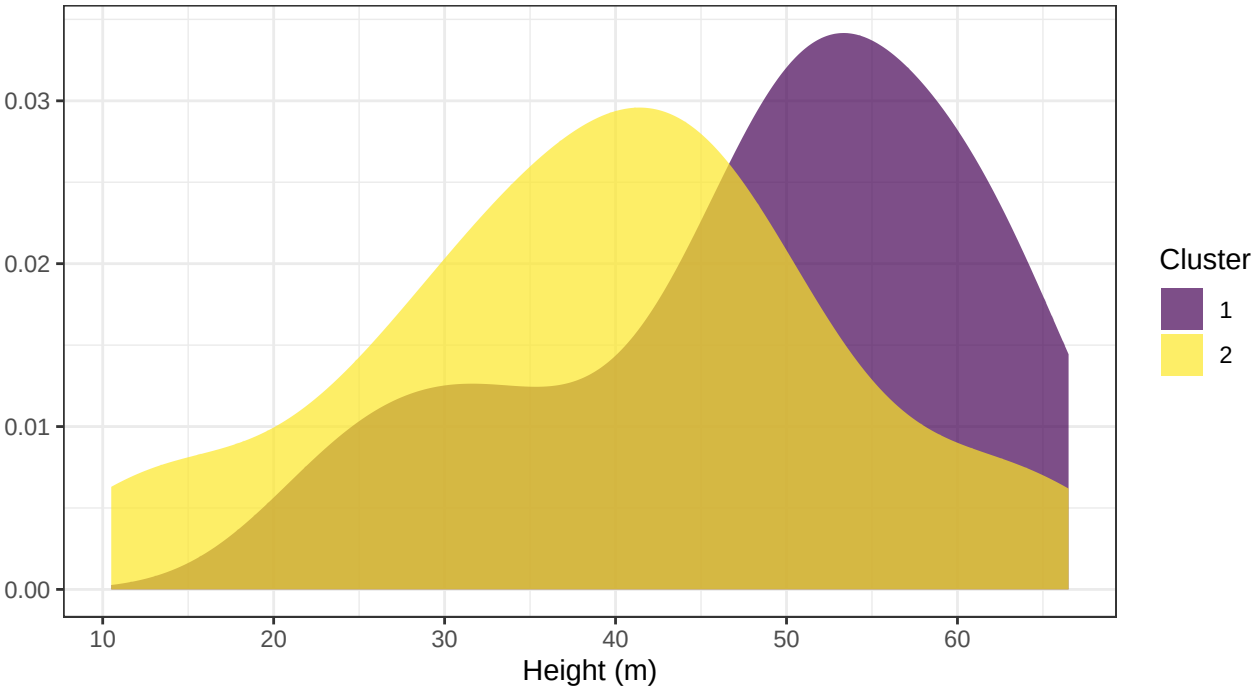

Figure 10: Distribution of height by cluster.

5 Graph Critique

Because of the heavy-tailed distributions of reflected and incident PAR, a log transformation yields a better visualization, as seen in figure 11. An important detail in this figure is that we have many values of reflected and incident PAR that are zero, so taking the log transform removes these values. This may still be useful for analysis that do not care about zero-valued readings.

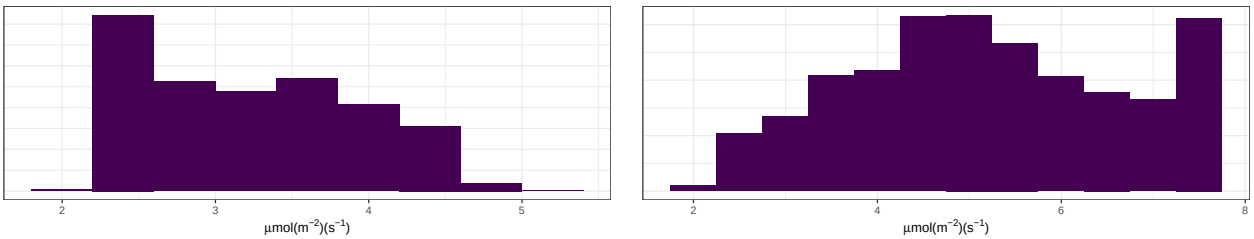

Figure 11: Distribution of log-transformed reflected PAR (right) and incident PAR (left).

The figures 3[c] and 3[d] in the paper by Tolle et al. (2005) try to showcase the relationship between height and temperature, humidity, reflected PAR, and incident PAR. Even though we think that the plots do a good work showing what the authors intended, we also think that there could be better ways to show these findings. First of all, a better visualization is given by the time series in figure 8. Another useful visualization for this relationship is the one shown on figure 10. Finally, another visualization could be made by creating height bins instead of showing boxplots for every height value. We create three bins, labeled "low," "medium," and "high" by dividing height in three intervals of the same length. One example of the resulting visualizations is shown in figure 12. Even though the distinction between bins is not perfect, this graph is easier to read than the ones in the paper and tells a similar story.

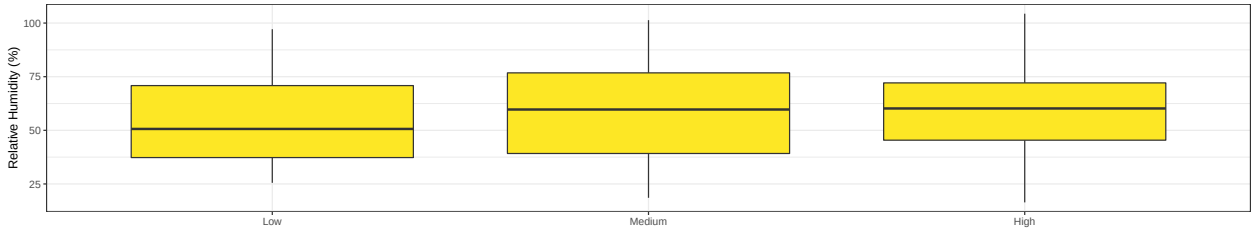

Figure 12: Distribution of humidity by height group.

For the first two plots in figure 4 in the paper, the colors are difficult to tell apart and do not convey any meaning. Also, the plots on the right are difficult to interpret. A better visualization is given by the time series in figure 8. By choosing an appropriate color gradient, we can mix the information shown in the two original plots in a single time series visualization.

Lastly, in figure 7 of the paper the authors try to show the differences between the log and network yields. We think this visualization could be more clear if the authors plotted both data sources in the same graphs, putting the bars side by side and using color to distinguish between the log and the network.

Bibliography

MPR/MIB User's Manual. 2004. 41 Daggett Dr., San Jose, CA 95134: Crossbow Technology, Inc.

Tolle, Gilman, Joseph Polastre, Robert Szewczyk, David Culler, Neil Turner, Kevin Tu, Stephen Burgess, et al. 2005. "A Macroscope in the Redwoods." In *Proceedings of the 3rd International Conference on Embedded Networked Sensor Systems*, 51–63. SenSys '05. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1098918.1098925.