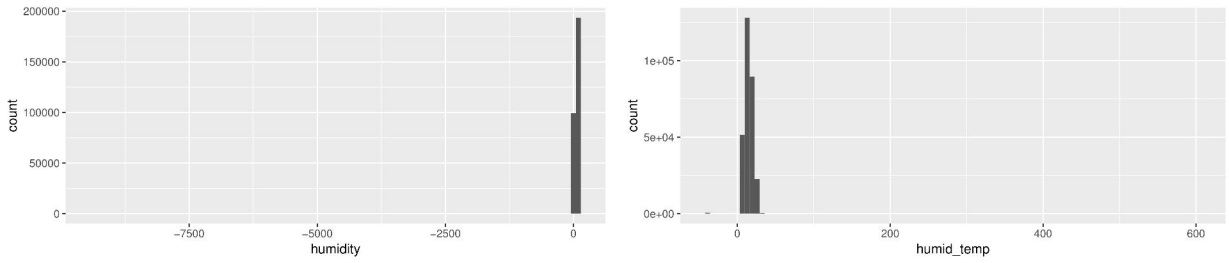
1 Data Collection

The purpose of this study is to analyze the microclimate surrounding a 70-meter tall redwood tree by using wireless sensors to record the variations in spatial (space) and temporal (time) dynamics, considering characteristics such temperature, humidity, as well as both direct and reflective solar radiation (light). These photosynthetically active radiations, PARs for short, are ultimately able to help scientists understand important biological and environmental impacts, like the amount of energy available for photosynthetic activities and carbon balance in the forest. Things like temperature and humidity grants scientists a deeper understanding on concepts such as the flow of sap or water through the tree. In this case study, environmental changes and patterns around the redwood tree were recorded from April 27, 2004 to June 10, 2004, a total of 44 days. The early summer period was chosen for this study to maximize the variations observed. The main variables of interest in this study are temperature, humidity, the amount of incident or direct radiation, and reflective or indirect radiation. The data collection process for these variables is operated by a network of sensors called macroscopes, a technology that must be protected from the weather while at the same time, exposed to track the microclimatic changes. First, these sensor nodes are thoroughly calibrated and then scattered around the tree vertically and horizontally. Vertically, they are placed both near the base and tip, from 15 to 70 meters above the ground with a 2-meter spacing. Horizontally, they are placed 0.1 to 1.0 meters across the trunk's radius. Nodes can also be described by their direction, whether they face the east or the west. The substantial variations one observes create gradients that are observable through plotting and performing explanatory analysis. In addition, readings from the sensors are taken once every five minutes, equivalent to one epoch, and the data proceeds to be transmitted to a database. The data in sonoma-data-log.csv contains data taken from the logging itself and the separate network file is there to compensate for the loss data in the loggings in the case of unexpected failures in battery and such. Conversely, the logging file can do the same for the network. Through inspection, one notices that the data collection process differs between the two files, and these differences should be taken into consideration.

2 Data Cleaning

2.1 Histograms

Upon intial inspection of the histograms, it was clear that the presence of outliers would make analysing the distributions a difficult task. For example, consider the histograms from the sonomadatalog file. See figure 1



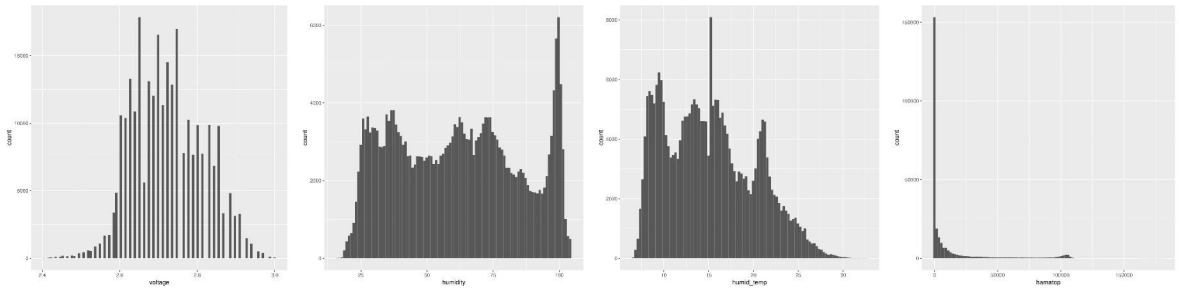
(a) Humidity

(b) Temperaturel

Figure 1: Histograms

Extreme outliers mean the range that must be captured to include all datapoints is immense, and it is difficult to examine closely the relevant pertubations in each variable. This is a symptom of all variables in the sonomadatalog file. Now, ofcourse we could limit the domain, effectively diregarding the outliers, and allowing us to focus on the relevant data. However, the question arises, where do we draw this line?

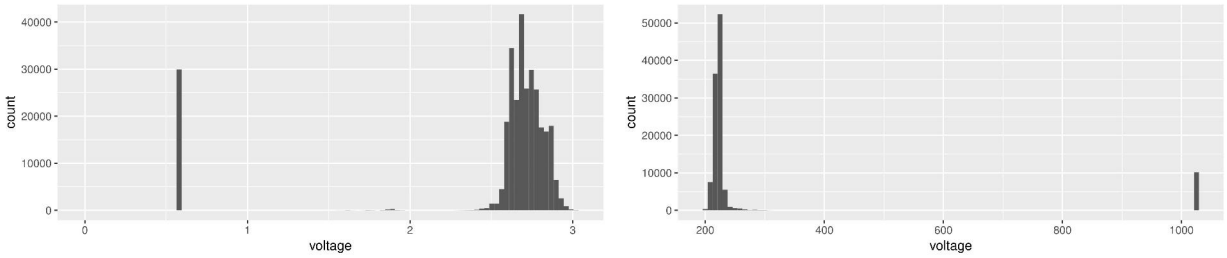
Decisions based on a visual inspection should be made with extreme caution.We will rely on the author's outlier detection methods as a basis for the more informed outlier removal methods we will gradually develop in this section. Indeed, once a node's battery voltage strays from a healthy 2.4 volt - 3 volt range, the data of the node is considered inaccurate. As noted by Tolle, excluding faulty nodes managed to eliminate nearly all outliers. We will adopt this appoach to attain our histograms. See figure 2



(a) Voltage (b) Humidity (c) Temperature (d) Incident PAR

Figure 2: Edit Histograms

The next problem arises when we consider consistency between the datalog and datanet files. See figure 3



(a) Voltage $data_log$ (b) Voltage $data_net$

Figure 3: Voltage Range Inconsistency

While datalog reports voltage in volts (analog signal), datanet uses a digital form. To coerce the voltage data in datanet to the same range, we must convert the data from digital to analog form. However, this is not a typical ADC. As can be seen in figure 2, where large values of voltage in raw datalog file correspond to small values in raw datanet file. Typically, large analog values would correspond to large digital values. To fix this, we will return to the original files, and compare outliers. It is reasonable to assume the prominent outlier in datalog is associated with the prominent outlier in datanet. Proposing an inverse relationship between the two measures:

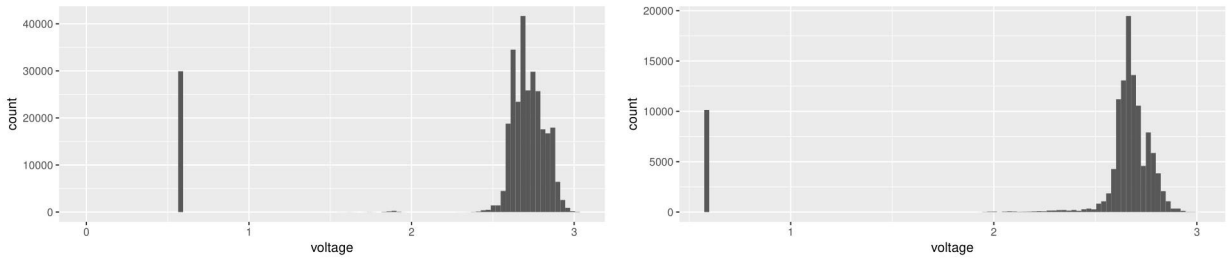
$$data_log = \frac{a}{data_net}$$

We calculate the constant, a, and convert the voltage measures of the troublesome data file. See figure 4

Now we can run the same voltage based outlier rejection strategy for the datanet file, and construct histograms. It is also important to clean repeats in our concatenated data file.

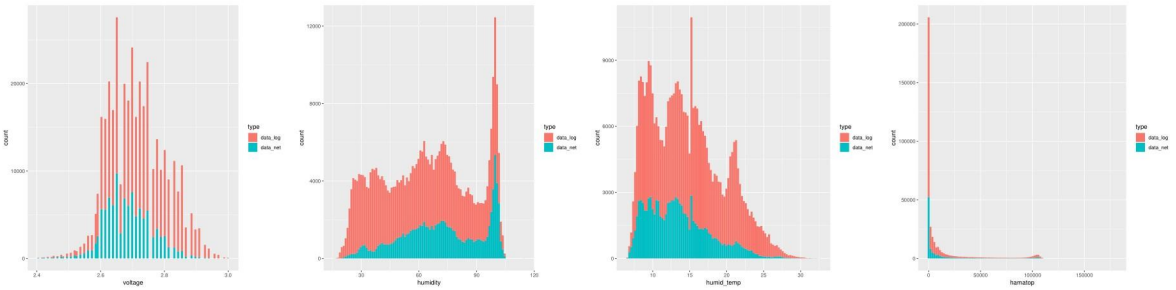
2.2 Remove Missing Data

There are 10059 missing values. They all appear (every day) between May 7th and May 26th. During the day, errors appear quite regularly, every 3 or so minutes. However, there are days during this time period



(a) Voltage $data_log$ (b) Voltage $data_net$

Figure 4: Voltage Range Consistent



(a) Voltage (b) Humidity (c) Temperature (d) Incident PAR

Figure 5: Combined File Histograms

where errors are more frequent. 24 th of May being when the most errors occurred, and the 7th of May is when the least errors occurred.

2.3 Main Table

Nodeid was used as a key to find Height, Direction, Distance and Tree values to incorporate inot our main table. There are 15 variables in the new dataframe

2.4 Outlier Search

Outlier search is conducted using our main file: a refined concatenation of datanet and datalog files. THe premise for our outlier search is two part. A histogram provides an overview of the data and its distribution, revealing any overtly outlying points. A boxplot will refine the search and points lying well outside the maximum and minimum bounds will typically be considered outliers. The bounds are defined:

$$IQR + -1.5 * IQR$$

It is important to critique the almost callous techniques used to eliminate outlilers in this section. While it is possible to justify removing datapoints based on deviation from mean from a purely statistical viewpoint, we must take into consideration the broader context at hand. What matters is not how far a data point is from the mean or from its predicted value. What matters is how much that one data-point influences the conclusions that will be drawn from the analysis. Outliers may be the only data telling you the true story. The author justifies removing extreme values based on the faults in the deployed sensors. We use these values as a final error check, but also to justify the deviation based outlier elimination employed

3 Data Exploration

3.1 Pairwise Scatterplots

We know missing values appear starting on May 7th, and as noted by Tolle, most nodes seem to 'die' on May 26th. Therefore the time period we select for this analysis should be before May26th, however we must reconcile the fact that the presence of errors (missing values) even before this time mean an absence of data throughout the entire spectra of time. We choose a week where missing values appear low : from May 10th to May 17th. To find this data we must rely on the epoch data, and the fact that the ???rst reading was taken on Tuesday, April 27th 2004, at 5:10pm, and our last reading was taken on Thursday, June 10th 2004, at 2:00pm, nearly 44 days later. We will base all time sensitive analysis on the epoch readings to avoid the trouble of converting the faulty time values given in the datalog file

Temperature vs Humidity:There appears to be a linear relationship between humidity and temperature. Specifically increases in temperature correlate with decrease in humidity. High temperatures would cause evaporation of moisture in the air , and thus we would expect to see humidity decrease. *Voltage vs Temperature*:A clear positive and linear correlation between temperature and voltage is evident. While it has been established that extreme voltage values produce erroneous detections of temperature, an explanation for the strong linear link between the two variables within reasonable bounds is glossed over in the paper. *Height vs Incident PAR*:We expect to see Incident PAR as being higher at higher heights. However, from the simple scatterplot construcuted it is difficult to see any such relationship. A more precise and refined dataset must be used to search for any such relationship. For example, perhaps we should exclude all measurements made at night, or perhaps we could construct a time series; analysing how PAR travels through the height dimension over the course of a day. *Temperature vs Incident PAR*:We expect to see increases in temperature with increasing incident PAR. Theoretically, sunlight should go hand in hand with temperature, however,this plot does not clarify any such relationship. This can be explained by the position of nodes, both in terms of height and also direction. Detectors at higher levels, unencumbered by foliage and unobstructed are likely to exerience the full dynamism of sunlight thorughout a day. Conversely, detectors at lower levels may never detect significant Incident PAR, although they may still experience a full range of temperature as sunlight filters through the canopy heating the lower forest areas. This may explain why even small "hamatop" values are associated with the full spectrum of temperature

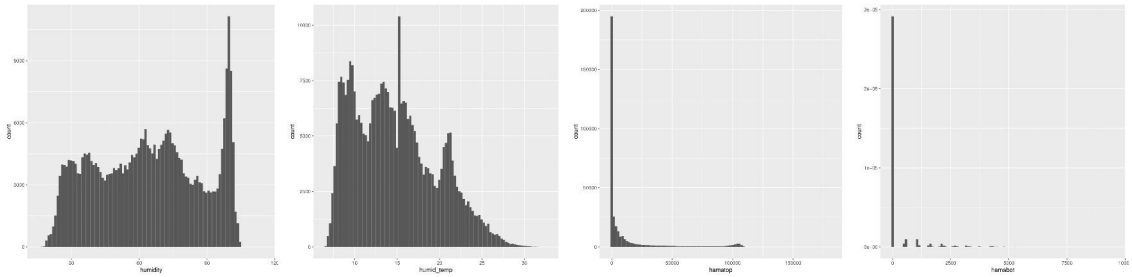
3.2 Association with Incident PAR

Here we will attempt to eliminate some of the obscurity of data we discussed in the previous section. By examining the factors influencing Incident PAR measurements and clarifying the nature of any relationships we find.

A notable characteristic of both plots is that the density of points with lower humidity decreases as we move to the upper echelons of Incident PAR measurement. The opposite is true for temperature. Higher levels of Photosynthetically Active Radiation indicate larger levels of sunlight saturation, and we expect the prevelance of sunlight to have a diminishing effect on heat dependent variables such as humidity. Plots suggest that the variability of temperatures and humidity levels is reduced at locations exposed to large levels of Incident PAR, and more specifically, that these locations are limited to a higher band of temperature and humidity conditions. What is strange is that high temperatures are less common with high radiation, contrary to what we expect. This motivates our next move, to understand the behaviour of Incident PAR across the time and height dimensions. It is baseless to understand the relationships Incident PAR has with other variables if we do not also comprehend the characteristics of Incident PAR on its own.

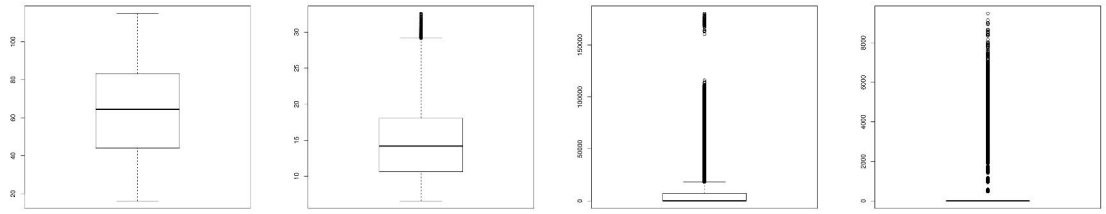
3.3 Temporal Trends

There seems to be a discontinuity in the humidity plot on day 7. Humidity plot and temperature plot appears to mirror each other, in that when we see increases in temperature, humidity decreases. Temperature ranges from around 5C to 24C. Humidity ranges from 30 to 92



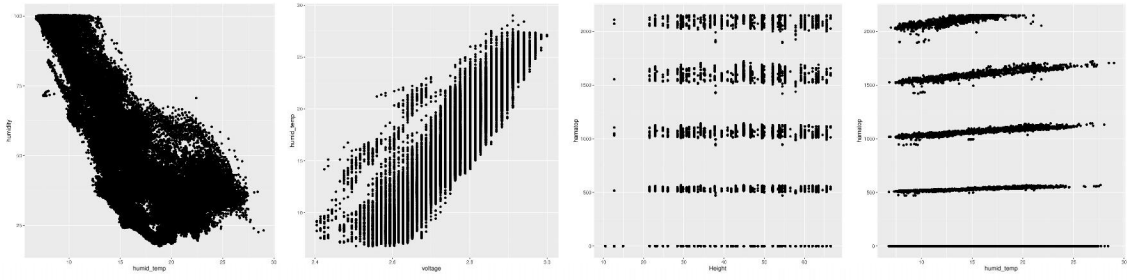
(a) Humidity (b) Humid$_t emp$ (c) Hamatop (d) Hamabot

Figure 6: Histograms



(a) Humidity (b) Humid$_t emp$ (c) Hamatop (d) Hamabot

Figure 7: Boxplots



(a) Temperature vs Humidity (b) Humid$_t empvsVoltage$ (c) Height vs Hamatop (d) Hamabot vs Voltage

Figure 8: Pairwise Scatterplots

3.4 PCA

```
## Standard deviations (1, .., p=4):
## [1] 1.3621590 1.0183058 0.9828439 0.3762899
##
## Rotation (n x k) = (4 x 4):
##                    PC1         PC2         PC3          PC4
## humid_temp  0.70536550 -0.06501571 -0.04794422  0.704225683
## humid_adj  -0.70754803 -0.01266172 -0.01744908  0.706336322
## hamatop    -0.04164538 -0.71921138 -0.68983102 -0.071650670
## hamabot    -0.01004767 -0.69162680  0.72217037 -0.004622705
```

As we can see, the first principal component is an equal mixture of humidity, temperature and adjusted humidity; although temperature is negatively correlated with the other two (and we expect this from the earlier histograms in this paper). The second component is almost strictly influenced by Incident PAR. We have defined a useful set of axes for a low dimensional projection of the data. The usefulness derives from the fact that we can understand the datapoints in context even after projection. Movement along the horizontal axis represents a direct compromise between humidity and temperature, while the vertical axis closely approximates Incident PAR.

4 Interesting Findings

4.1 Clusters on PCA

Both EM and K-Means choose the same clusters to divide the post-PCA data. This holds true over numerous iterations, and was confirmed by comparing group classification. K-Means divides data based on distance similarity arguments, while EM creates group based on variance calculations and by considering the underlying gaussian mixture model. The alginment of the two methods so closely idnicates that there is great internal cohesion within the three groups, and their separation, although not visually obvious, is significant. When we consider that the vertical PCA axis is a measure of radiation, and the horizontal axis a measure of humidity and temperature we can crudely attempt to classify overarching group strucutres within the data. Remebering the PCA components, the graphs show separation of data into these categories:High Radiation, Low Radiation-High Humidity-Low Temperature, Low Radiation-Low Humidity - High Temperature.Further analysis would perhaps expose that a variable such as voltage or node height may be influencing this consistent categorisation.

4.2 Trends in Temperature

Here we use epoch as determinant for date:time to bypass the faulty measurements given in the $data_{l}og file$.

The "Fixed Heights" graph shows the evolution of temperature at 3 height categories from 5.30pm to 5.30pm the following day. While at all heights, a similar trend is repeated. The evolution of temeprature on large and small timescales is evident

5 Graph Critique in the paper

5.1 A Better Histogram

The incident and reflected PAR both have long tails, so we propose that the histogram should undergo a log transformation in order to display the tail values more clearly. Ofcourse we run into the problem that the majority of Incident PAR and Reflected PAR data has value 0. We overcome this by shifting all data points by one, after converting them to the right units. See figure 17 and 18

5.2 The Problem with Plots...

The problem with figure 3.c) lies in the fact that the author used the overall dataset to construct boxplots, hoping to notice variation in variables with node height. The flaw in this approach is that broad averages of value-height data pairs need to be taken across the entire the entire lifespan of the experiment. Consequently, any relationships between variable and node height across smaller timescales, such as over a day, or over a week, disappear amidst the non-selectgive averages that are taken to construct the boxplots. To improve 3.c) perhaps it is better to look at time variation across a smaller time period. Alternatively, the author could select a moment during the day, say midday, and sample only "midday" data from the larger dataframe. See figure 19

Ofcourse, a problem with using data from a single day is that our data cannot be taken as being indicative of a "typical" day. While taking the whole dataset and forcing broad, spectrum-wide averages is inherently a better measure of dataset-wide patterns, the information it provides may not be useful. It oversees nuances in variable behaviour, such as how the distribution is affected by time of day. Perhaps if we want to understand typical behaviours, as the author does, we should fix the time of day, and then examine the whole dataset, in an attempt to minimise room for movement in variables (such as time) that are not the focus of the plot. See figure 20

Clearly there is a large amount of variance between different height points for temperature and humidity. PAR plots, being largely composed of zeroes, was not informative at all. Indeed a time series where we track the temperature gradient across time may prove to be the most imformatitve. This was done earlier in the report

In figure 3d) the author has shifted the mean of each variable to 0, highlighting deviation of values from their mean. The usefulness of this approach is demonstrated in the PAR plots. Almost immediately we identify that at lower forest heights, we detect below average radiation, as of course we would expect a thick canopy and foliage to obscure incident or reflected light. However, these patterns may not be the underlying structure of the distribution at all times of day. Obviously, at night time we would not expect to see this difference between the higher nodes and the lower nodes. There is no light! We do not fully adopt the centering technique used by the author as it is unclear. See figure 21

```
## Error in sonoma_data_night$hamatop - mode(sonoma_data_author$hamatop):  non-numeric argument
to binary operator
```

However, we do see that the average value at night, at each height, tends to be below zero. Zero is defined as the overall mean (day and night) specific to that height. As we expected, the distribution deviates to the left, lying largely below the overall average radiation value at each node height.

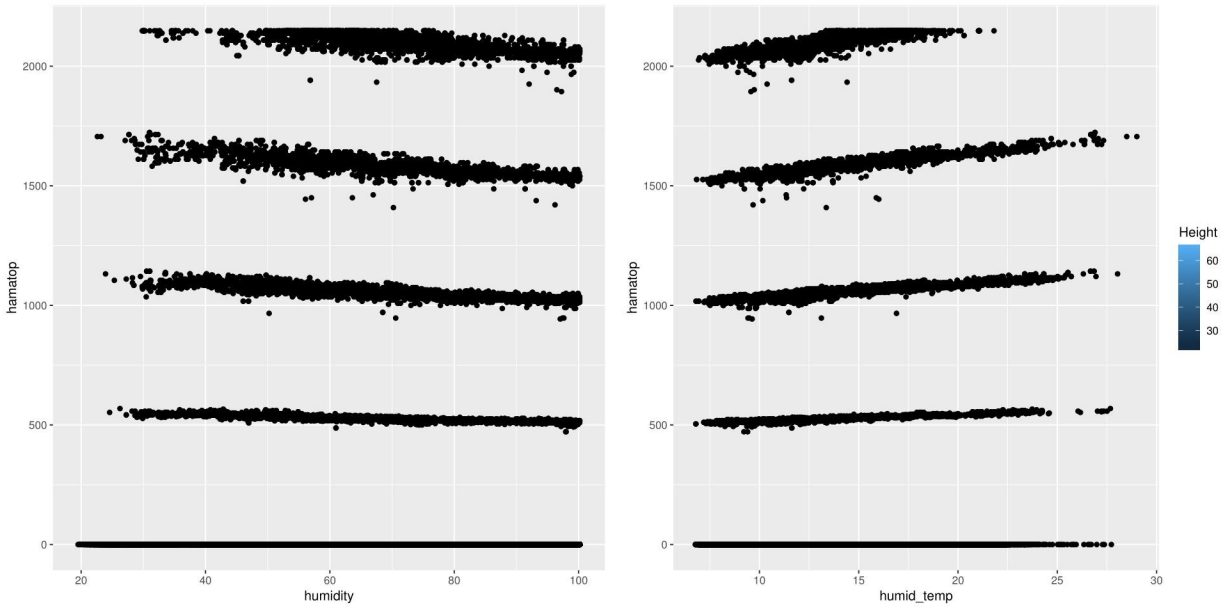
5.3 Plots in Figure 4

It is definitely difficult to distinguish all the colours, hence all the nodes, depicted in the plots. There are several alternatives to reduce the amount of nodes while still retaining a large portion of the information in the plot. One way is to naively and randomly sample a subset of the nodes to graph. A better way is to categorise the nodes into subsections, for example, by height. See our plot in the "Trends in Temperature" subsection for an example. See figure 22

5.4 Figure 7

While each plot presents a new perspective on the yield problem, it is questionable whether it is necessary to include some of these plots. The third plot, percentage yield vs node height, is not only difficult to read but seems redundant next to the adjacent plots. Plot 4 identifies the days where nodes at specific heights stopped working, and by visually judging the total length of the blue bars we get an idea of the total yield at every height. Ofcourse it isn't as accurate as plot 3, but an important aspect of plotting is producing a visual that is easy for a reader to understand. Plot 2 further specifies yield on every day, thus the necessity of plot

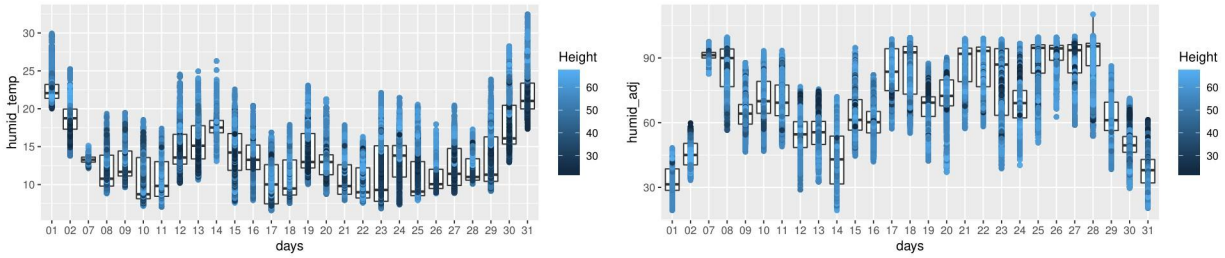
3 must be questioned. For each mote at each timepoint it was considered to 'report a 1' if it reports any data at all. However, constructing yield plots on this premise prevents exploration into what components of the sensor were faulty. Was it the temperature detection? Humidity? Radiation? Indeed, if we welcome this variation we may be able to identify what faults caused the differences in yeild in the network compared to the local log. Surely, it would not have been the same faults.



(a) Incident PAR vs Humidity

(b) Incident$_{P}ARvsHumid_{t}emp$

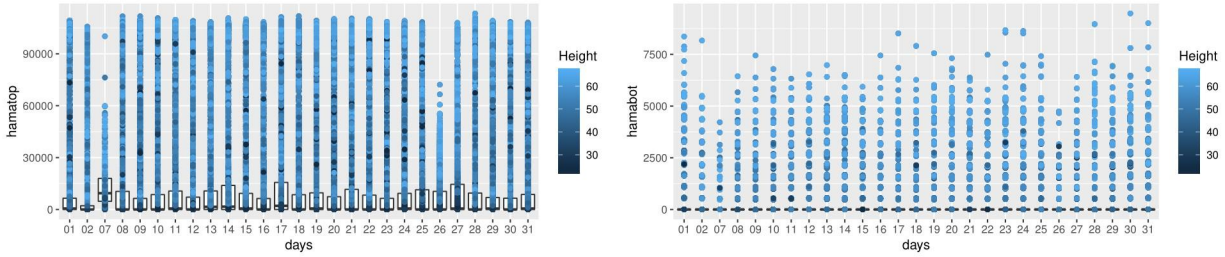
Figure 9: Incident PAR Relationships



(a) Temperature

(b) Humidity Adjusted

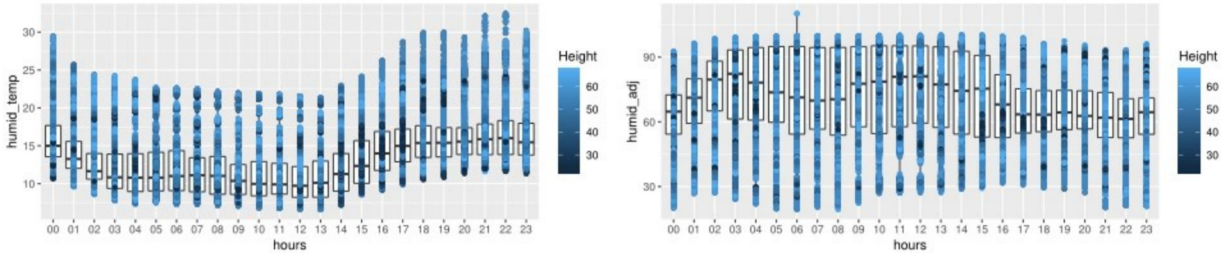
Figure 10: Evolution Over Time (Days)



(a) Incident PAR

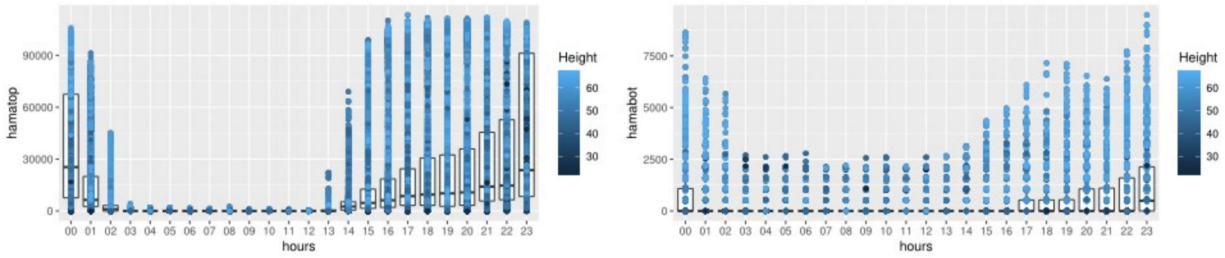
(b) Reflected PAR

Figure 11: Evolution Over Time(Days)



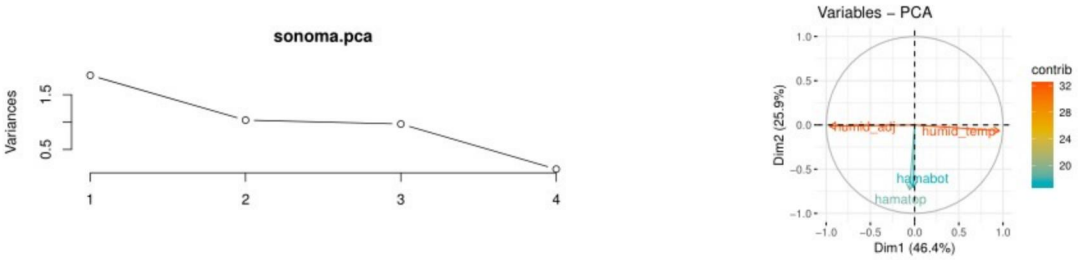
(a) Temperature (b) Humidity Adjusted

Figure 12: Evolution Over Time (Hours)



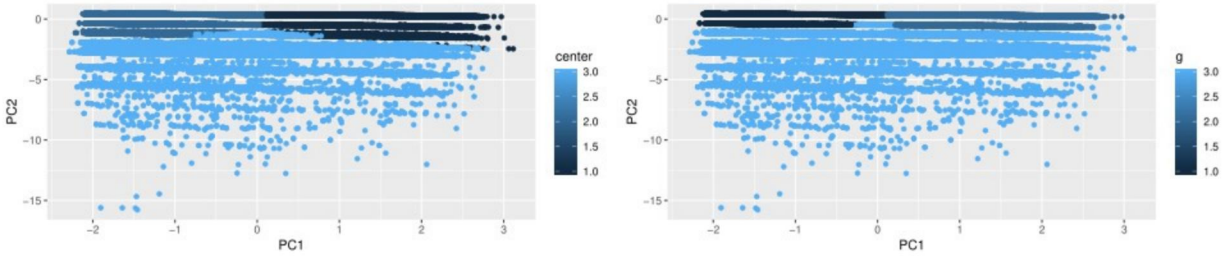
(a) Incident PAR (b) Reflected PAR

Figure 13: Evolution Over Time(Hours)



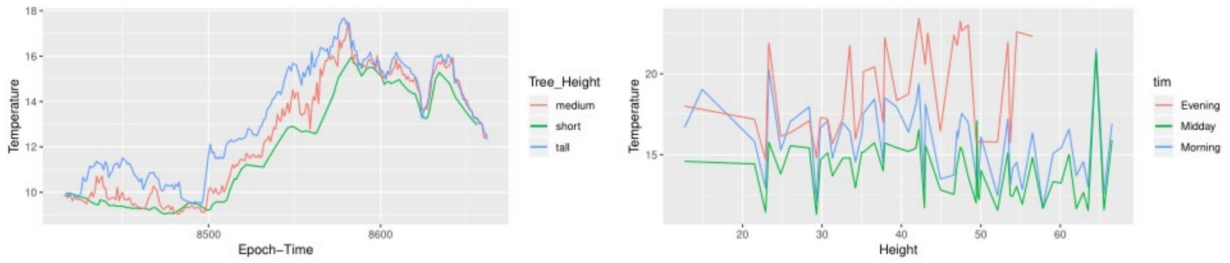
(a) Screeplot, Components (b) Screeplot, Components

Figure 14: PCA



(a) KMeans (b) EM Cluster

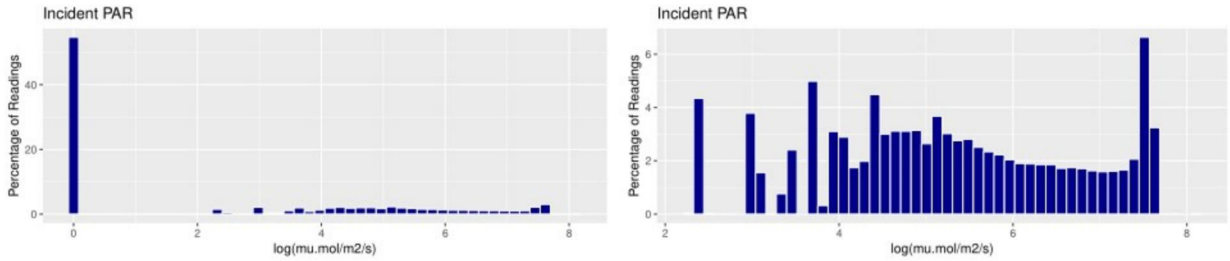
Figure 15: Height and Data



(a) Fixed Heights, Across A Single Day

(b) Fixed Times, Across Range of Heights

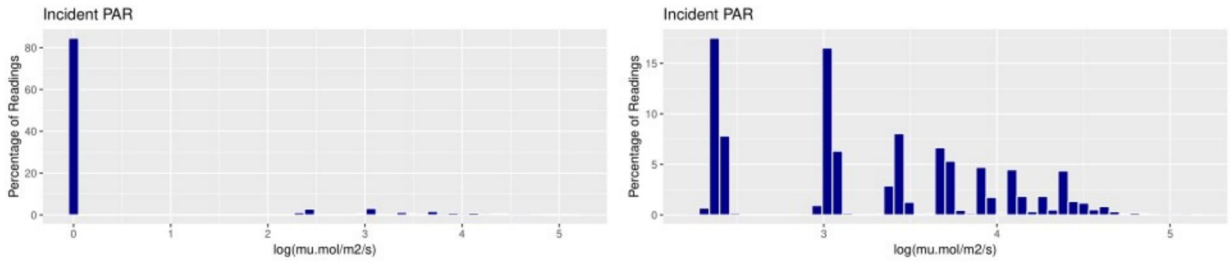
Figure 16: Trends in Temperature



(a) Log Transformed

(b) Zoom in on tail

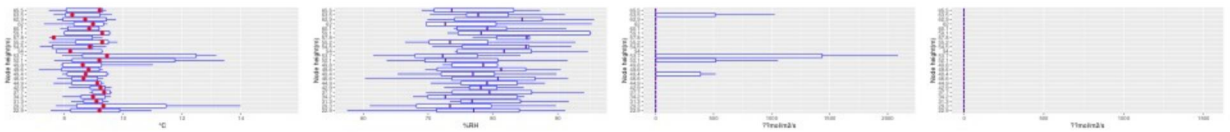
Figure 17: Log Transformed Incident PAR



(a) Log Transformed

(b) Zoom in on tail

Figure 18: Log Transformed Reflected PAR



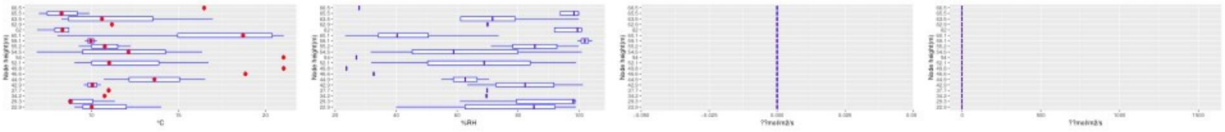
(a) Temperature

(b) Relative Humidity

(c) Incident PAR

(d) Reflected PAR

Figure 19: Variations and Height Over One Day



(a) Temperature (b) Relative Humidity (c) Incident PAR (d) Reflected PAR

Figure 20: Variations and Height During Midday, Over Experiment Duration

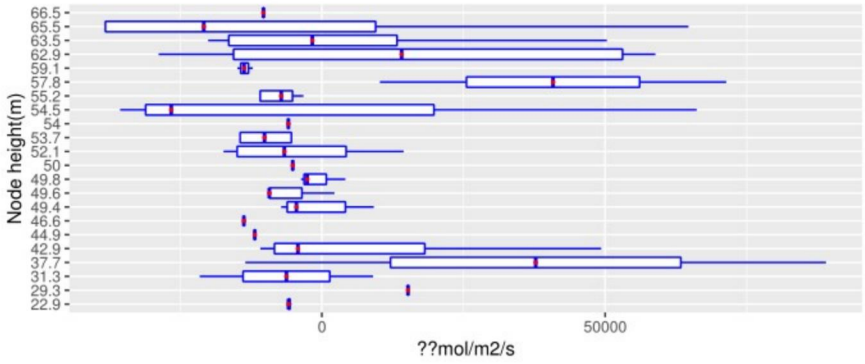


Figure 21: Centered Incident PAR

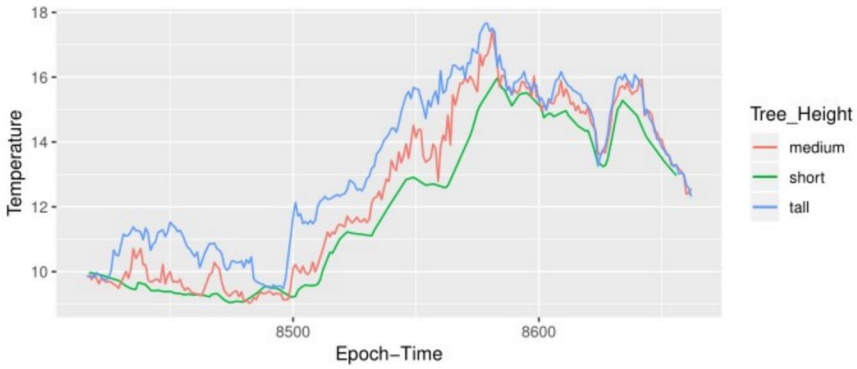


Figure 22: Trends in Temperature with Height