Project 1 Redwood Data Report

Phoebe Abramowitz(26386343) & Omri Newman(3032273024)

3/8/2019

Paper Summary (20 pts)

This study of microclimatic monitoring aims to get a better understanding of the life of a Redwood tree in Sonoma, CA. To do this, an interdisciplinary team from the University of California, Berkeley designed an experiment that paints a picture of a redwood tree's ecophysiology spanning fourty-four days in Spring 2004. They did this using a unique network of nodes and wireless sensors to gather spatio-temporal and environmental dynamic data over the entire organism. Each of the thirty-three nodes collected hundreds of thousands of data points, which lie in four different three-dimensional spaces (time x height x value). This wireless sensor network, or macroscope, allows for rich analysis on an unprecedented dataset.

After brainstorming with several biologists, Sonoma's coastal redwood forest was chosen as the site of this study because of its dynamic ecosystem. Longer daylight hours during the Spring provided sufficient time for the photosynthetically active radiation sensors to collect data, and the coastal fog coming off the Pacific guaranteed a wide range of temperature and humidity gradients.

A thorough analysis of the redwood data confirmed the biologists' hypothesis, which premised the existence of dynamic gradients present in the redwood trees ecophysiology. By projecting the data points onto a subset of the three-dimensions, the analysis task became simpler.

One drawback to using the sensor network was the presence of missing values throughout the data set in both time and height dimensions. However, the team worked in spite of this by studying distributions in most of their analyses. In addition to confirming the team's hypothesis, more information about the climatic distribution of the tree was extracted from the analysis. One impactful step was using the data to build a quantitative model of the effect of microclimatic gradients on the sap flow rate of redwood trees. By better understanding this process, biologists can piece together a more detailed picture of the larger scale carbon and water exchange within a forest ecosystem.

The main variables of interest in this study are those that shed light on the microclimate and ecophysiology of the redwood tree. Thus, each data point can be viewed in three dimensions– time, height, and value. The values, recorded by the network of sensors, are temperature, humidity, incident photosynthetically active solar radiation (PAR), and reflected PAR. Collecting this data over an extended period of time required a powerful system and precise deployment methodology. The data collection time period started in late April and spanned fourty-four days, sampling all sensors every five minutes. The nodes were placed along the seventy meter redwood tree starting at a height of fifteen meters, with two-meter spacing between nodes. The west side of the tree had a thicker canopy that served as protection against environmental effects, so the majority of the nodes were placed there. Lastly, the nodes were placed close to the trunk (0.1-1.0 meter) such that the recordings were of the microclimatic trends of the tree and not the wider climate.

Multiple sensors were integrated into the Mica2Dot wireless node, as a means to gather data on the environmental dynamics of the tree. This node has a one inch diameter form factor, an Atmel microcontroller running at 4 MHz, a 433 MHz radio from Chipcon operating at 40Kbps, and 512kB of flash memory. The node was packaged into a sealed cylindrical enclosure made from white high-density polyethylene that reflects radiated heat. The enclosure contained the node, battery, and two sensor boards - one for direct radiation and one for all other measurements. Data collected from these nodes were stored in a local database or gateway, then transmitted to an offsite database via a general packet radio service (GPRS) cellular modem. The UCB team used TASK software as the node operating system and data collection scheme to aid with this work. This multi-hop system utilized TinyOS and MintRoute systems to retrieve data from the nodes and store them in the gateway as efficiently as possible. The entire network of nodes was awake for four seconds every five minutes, in order to take sensor readings, transfer the data to the base stations, and then return to sleep. Two calibration schemes were completed by the team to ensure a satisfactory operation prior to setting up the final system. The first calibration consisted of leaving the sensors on a roof collecting PAR data every

thirty second for two weeks to verify the accuracy of the readings. The purpose of the second calibration was to understand the response of the temperature and humidity sensors. The nodes were tested in a controllable weather chamber that had varying temperature and humidity conditions, and recorded readings every thirty seconds. The calibration techniques used on the nodes took up space on the 512kB flash memory. Thus, some of the data loggers in the local database gateway ran out of memory before the end of the data collection time period. These loggers were meant to serve as a backup in case of network failure, and stored data in the case of lost GPRS connectivity. On the other hand, the logger data that was stored in the gateway could only be accessed at the end of the data collection time period, and there was lost information in the nodes that ran out of flash memory. This defines the main difference between sonoma-data-net.csv and sonoma-data-log.csv, which both contain values the other is missing.

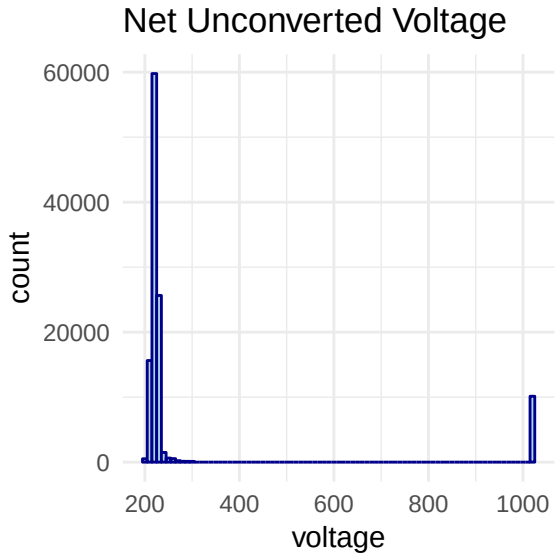
Data cleaning (40 pts)

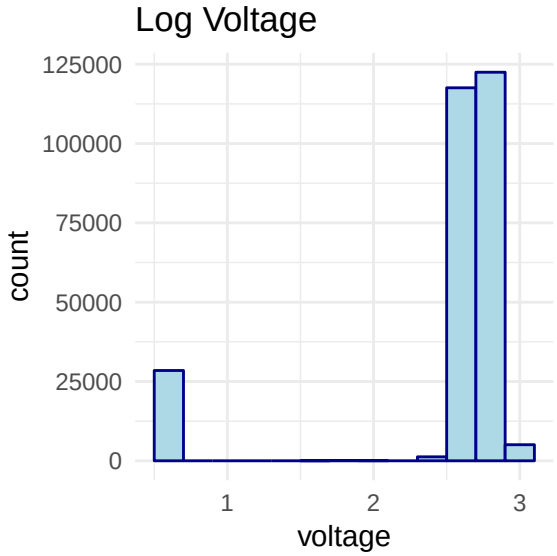
This data set contains gross outliers, inconsistencies, and lots of missing values. We use basic exploratory data analysis and data cleaning techniques to get a better understanding of the data, determine which values make sense in the context of the study, and remove missing values when necessary.

1. For the initial comparison in hamatop (Incident PAR) between **sonoma-data-log.csv** and **sonoma-data-net.csv**, we remove outliers that are ten times larger than the third quartile in an effort to explore the ranges more thoroughly. We notice a slight bump in the right tail of incident PARs histogram for **sonoma-data-net.csv**, which does not occur in **sonoma-data-log.csv**.

The histograms for hamabot (Reflected PAR) reveal some interesting trends shared by **sonoma-data-log.csv** and **sonoma-data-net.csv**. Not only are their ranges the same, but they both exhibit the same gaps in the values of hamabot that occur.

It is clear from the histograms below that voltage readings are not consistent between **sonoma-data-log.csv** and **sonoma-data-net.csv**





To reconcile the inconsistencies between voltage and PAR, we applied two separate conversions as to match the ranges from the paper:
The voltage readings in **sonoma-data-log.csv** are generally within the correct range of just a few volts, while the readings in **sonoma-data-net.csv** had significantly higher values. This discrepancy can be attributed to the analog-to-digital converter used within the ATmega128 node[1]. To switch from the 10-bit ADC readings to actual voltage measured, we followed a simple ratiometric conversion scheme [2]. We multiplied the values by 12.33 and divided by 1023 to transform the data into the desired range.

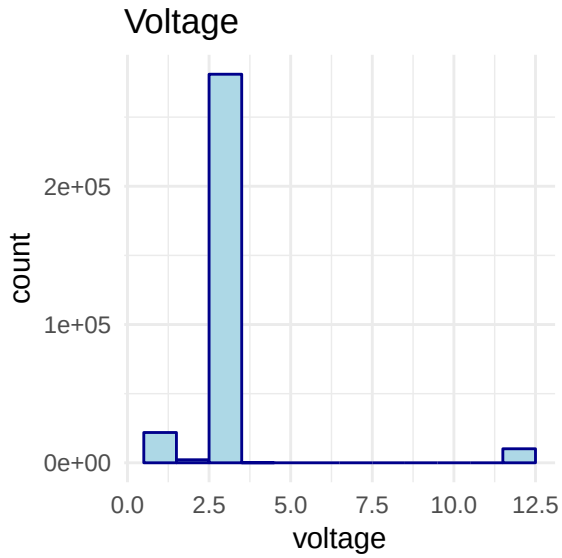
We also noticed the ranges for hamatop and hamabot differ from the paper. We convert the measurements of both types of PAR from Lux to PPMG, the units expressed in the visualizations in the study, using a 0.0185 conversion factor.

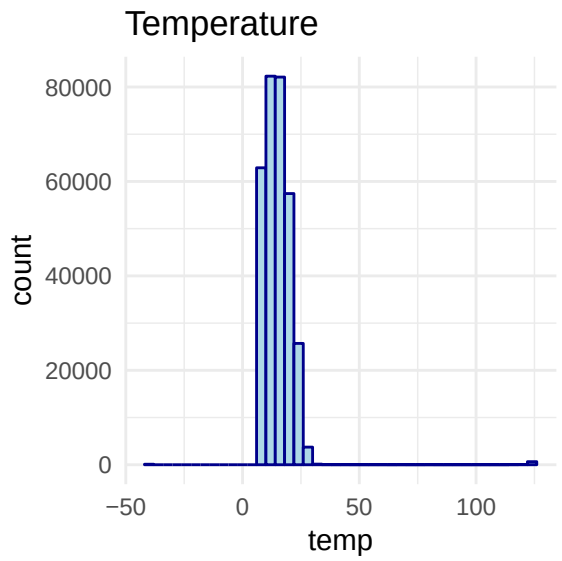
We then combine data sets from the log and the network into one main data set. We take the values from the network dataset in cases of dual instances, as the times are more precise.

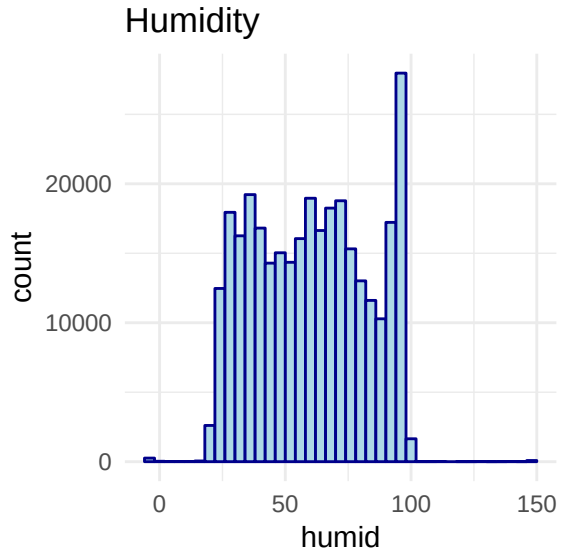
In our main data table, we remove 3686 rows where the temperature, humidity, reflective PAR, and incident PAR are missing. There are 3686 values missing for each of the 4 variables.

After removing missing data, reconciling inconsistencies, and joining the **sonoma-data-log.csv** and **sonoma-data-net.csv**, we incorporate the **mote-location-data.txt** using nodeid as a unique identifier for each sensor. All together we have the location of each mote within the two trees, along with the respective sensor data. We selected only the variables of interest amongst the 11 to create a desired dataset of all the readings. Our main table has 337,744 observations and 12 variables.

We chose to match the range of each variable as they are presented in the paper. It is clear in the following histograms that all of the variables contain some outliers. Instead of removing these data points per variable, we noticed that filtering out rows with faulty voltage readings–which are not reliable, as detailed in the paper– does the same job in one fell swoop. Just as is described in the paper, filtering out readings below 2.4V or above 3V removes 33,833 observations.


Voltage
count
2e+05
1e+05
0e+00
0.0
2.5
5.0
7.5
10.0
12.5
voltage


Temperature
count
80000
60000
40000
20000
0
−50
0
50
100
temp


Humidity
count
20000
10000
0
0
50
100
150
humid

(Bonus) Discuss other possible outliers and explain your reason why it is better to remove them than to keep them.

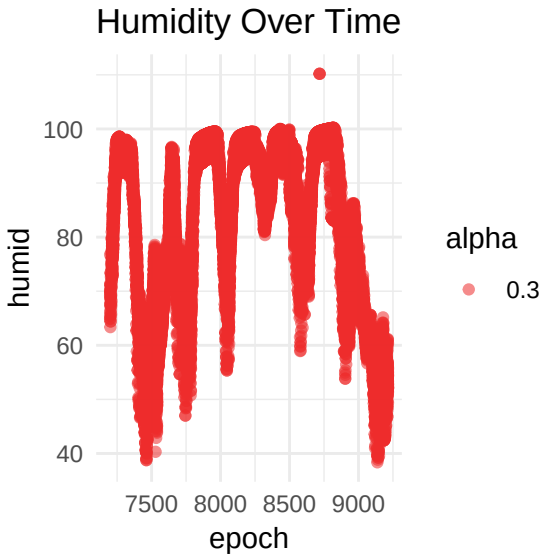
Outliers in this dataset are ones with recorded sensor values that don't fit the context of this study. For example, a temperature reading of -40 degrees celsius in the middle June is clearly a mistake. While most of these issues were the result of faulty voltage readings, others were not; these outliers can be explained by a number of reasons. The paper details a robust deployment methodology, which utlizies multi-hop routing between sensors and the local gateway. Some of these faulty readings can be attributed to issues with the GPRS as it was sent to the offsite database, or how the data was queried using TinySQL over the TinyDB and TASK software. The more variables that are introduced, the more room there is for error, especially when working with electrical systems in the middle of a redwood forest. It is certainly better to remove these outliers because they would add unnecessary noise and misleading values to the data, and subsequently to the data analysis. If biologists wish to create quantitative models of sap flow rate, then these outliers will disproportionately affect their model.

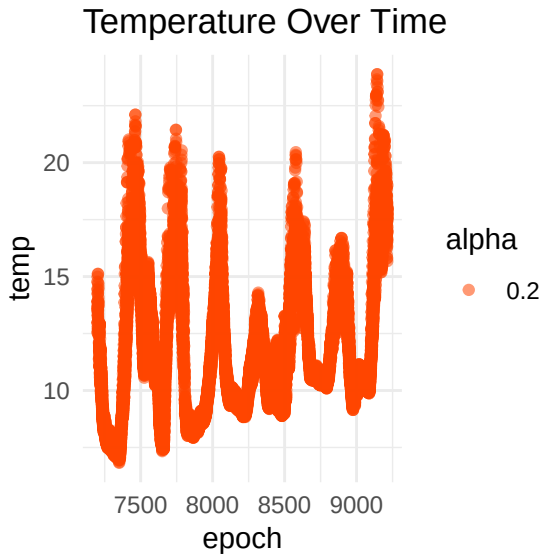
Data Exploration (40 pts)

- We choose the subset of the data occuring during the week of May 23rd through May 30th. There are more hours of sunlight in the last week of May as the summer solstice gets closer. This subset–containing ~28,000 values– in particular has fewer missing values than other time periods

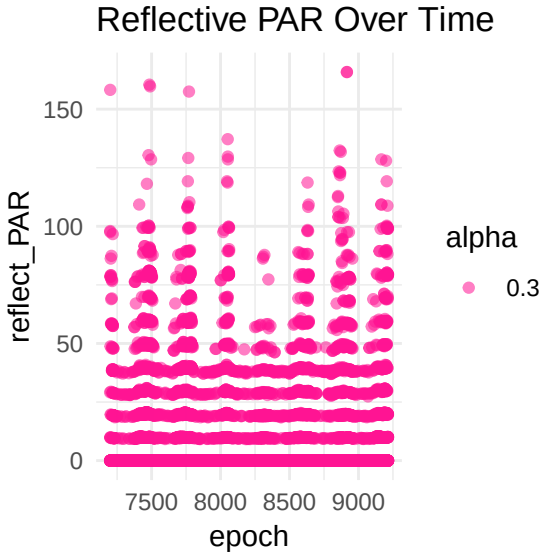
Some pairwise scatterplots of some variables and descriptions of our findings:

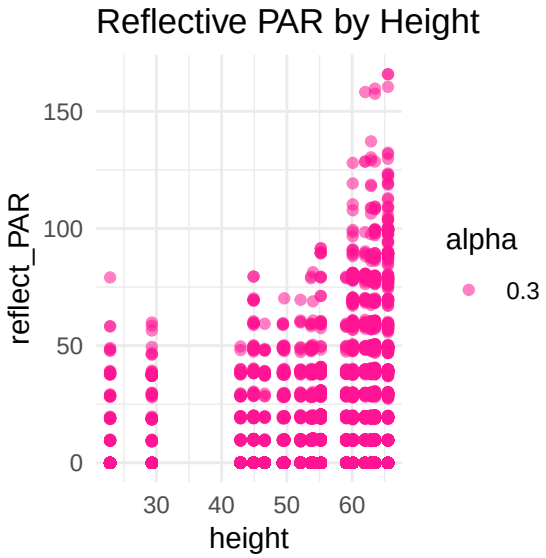
The number of the nodeid is not meaningfully correlated to height. Humidity and Temperature by time and height





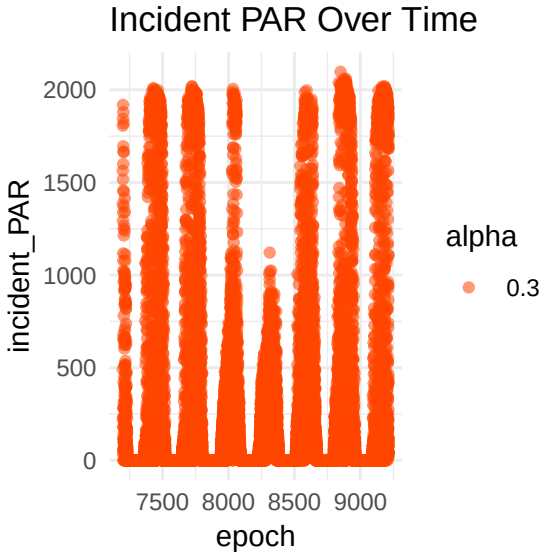
Reflective PAR over time loosely indicates cycles of daylight, and the maximum value observed of reflective PAR increases with height.

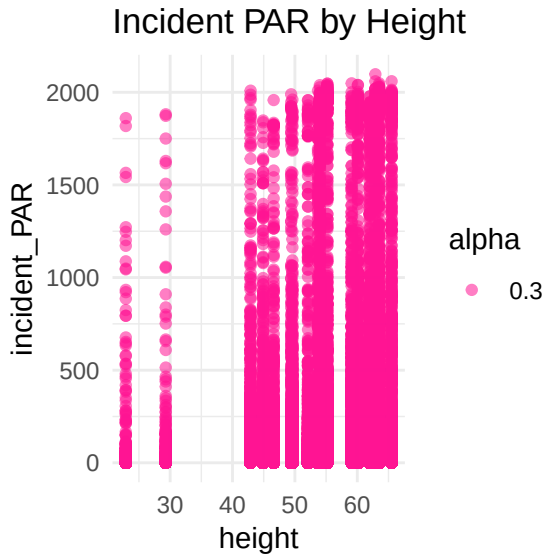


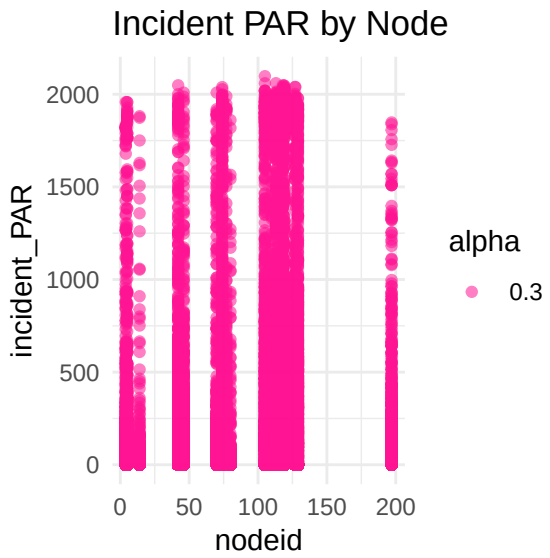


Are any of the predictors associated with Incident PAR? If so, explain the relationship.

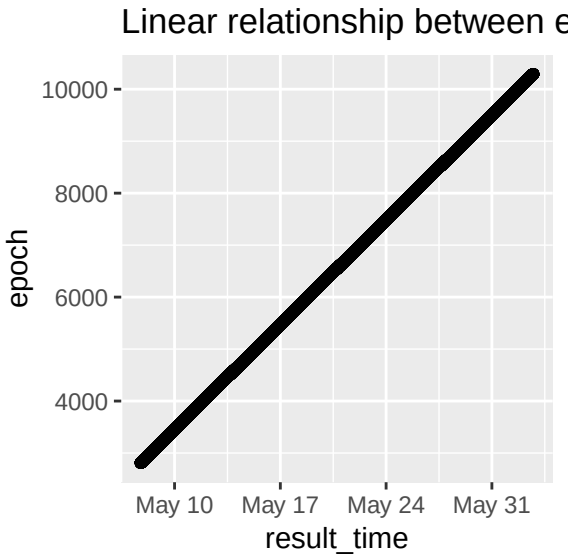
Incident PAR is visibly cyclical over time. The relationship can be explained by the daily daylight cycle of the sun rising and setting, since Incident PAR measures direct sunlight. The second and third plots here show that certain nodes are exposed to more direct Sunlight, but not necessarily by height



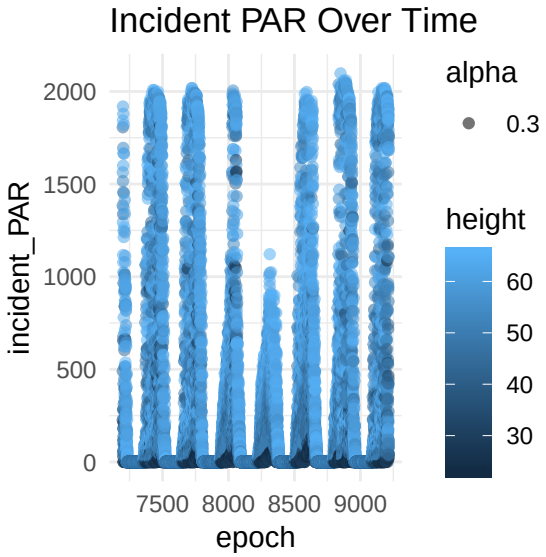


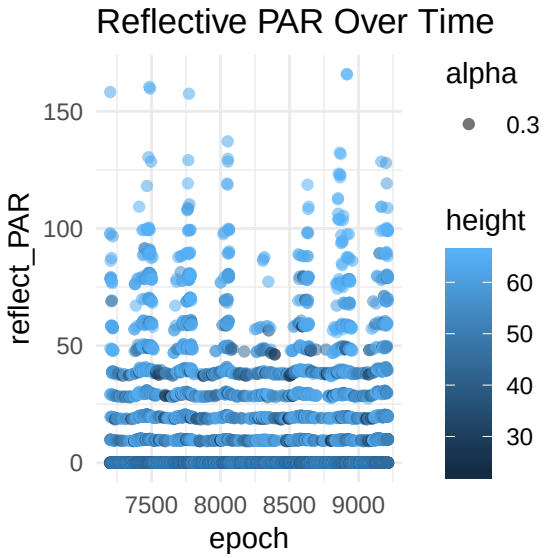


*Next, we consider each variable in the context of it's temporal trends, with height as color cue: (You can do it for different time scales during an hour, during a day or during the entire experiment). However, at least the plots with days as x-axis are required. As shown in the histogram below, I used epoch as the x variable for time here becuase their positive, linear relationship justifies our use of epoch as a standin for time that doesn't lose any information from data_log. We could also use result_time.

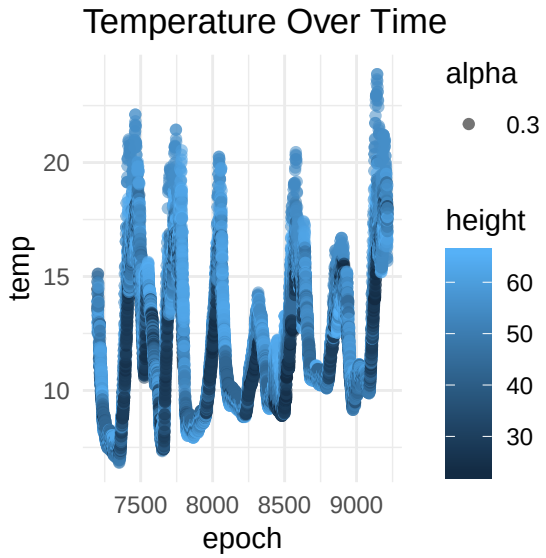


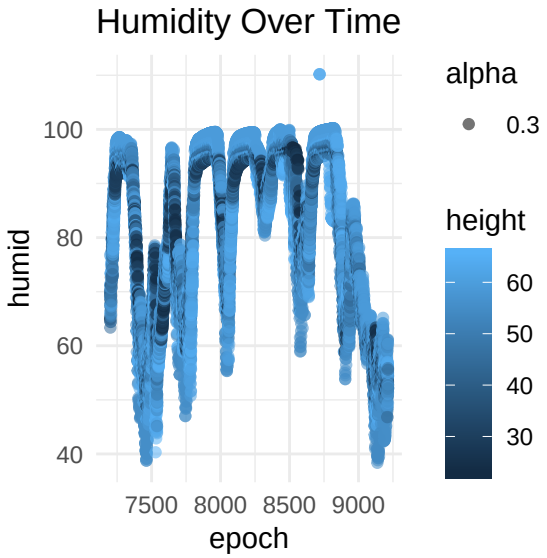
During the day, incident PAR is noteably least at the lowest height of the tree. However, there's not a height gradient throughout. Reflective PAR lacks this trait.





The lowest temperatures occur and the least height when and only when temperature is rising during the day.





PCA analysis

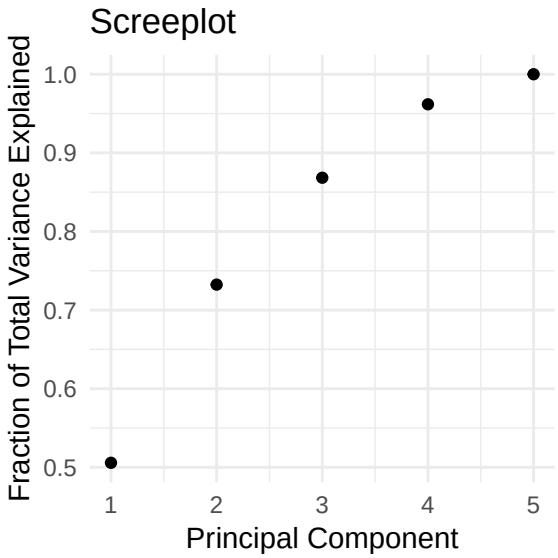
Can this data be approximated by some low-dimensional representation?

PCA analysis & scree plot of the data: Before running PCA on the desired subset of the data, we select five variables of interest (temperature, humidity, incident PAR, reflected PAR, height). We use Kaiser's criterion to choose principal components, of those with variances greater than one. We see out of the five eigenvalues, the first two are greater than one. The scree plot also shows that 73% of the variability in this data is encompassed by the first two principal components. For these reasons, we conclude there is a two-dimensional representation of this multi-dimensional data.

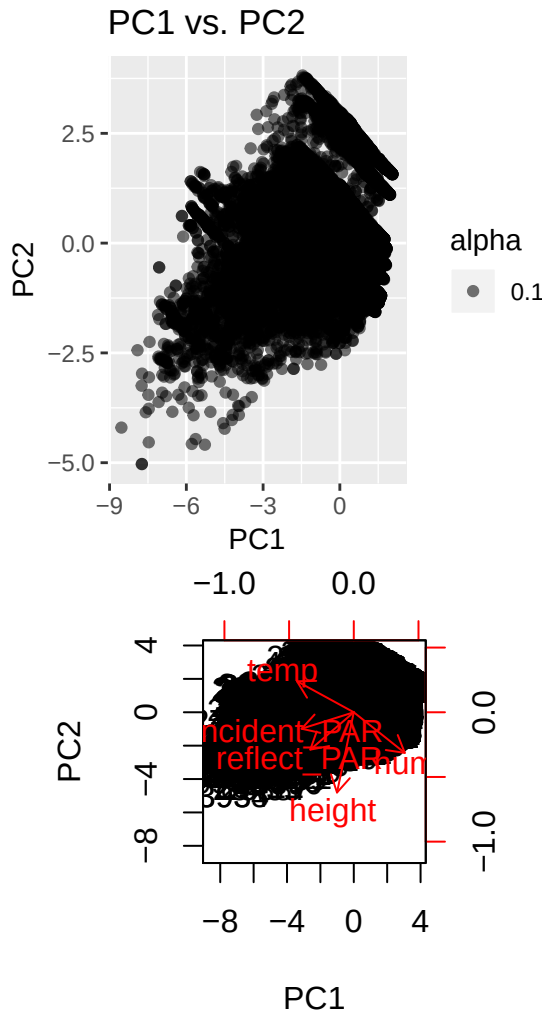
```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5
## Standard deviation     1.5902 1.0645 0.8246 0.68342 0.43707
## Proportion of Variance 0.5058 0.2266 0.1360 0.09341 0.03821
## Cumulative Proportion  0.5058 0.7324 0.8684 0.96179 1.00000
```

```
## [1] 2.5287710 1.1332366 0.6798909 0.4670686 0.1910329
```

```
## [1] 5
```



The following scatter plots show PC1 places a high value on temperature and both PAR's, while PC2 places a high value on height. Both PC1 and PC2 place roughly the same weight on humidity. The nodes which experience high temperatures and high PAR readings will have a greater first principal component in absolute value, as is seen in the biplot.

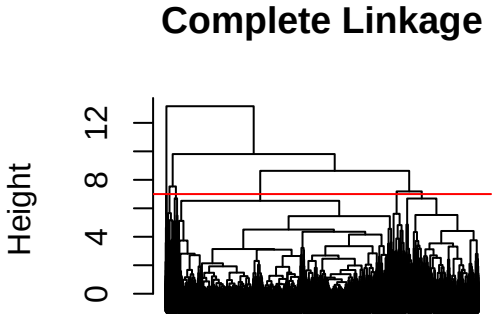


Describe two interesting findings from exploratory analysis of the data and comment on these findings.

Finding 1.

Unsupervised learning techniques like clustering allow us to study patterns within multi-dimensional data sets. To start off, we run hierarchical clustering with complete linkage on the subset of interest, and notice that six clusters is a good starting point for this data.

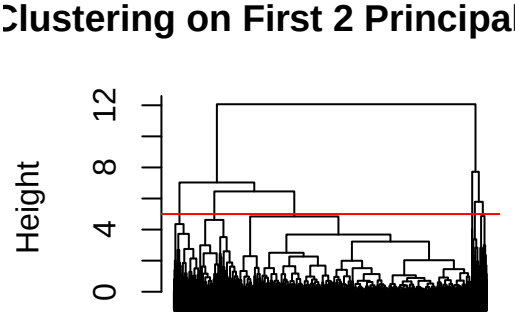
```
##             Length Class  Mode
## merge       55850  -none- numeric
## height      27925  -none- numeric
## order       27926  -none- numeric
## labels          0  -none- NULL
## method          1  -none- character
## call            3  -none- call
## dist.method     1  -none- character
```



We compare K-means clustering with K=6 to the hierarchical clustering that was just obtained, and make note of the differences. There are no zeros in the table below indicating that there is overlap between all clusters obtained by both K-means and hierarchical clustering.

```
##
## km.clusters     1     2     3     4     5     6
##           1   485   507     7    78     0     0
##           2   306     4   394    11   398   134
##           3     0     0  4072     0 14231    24
##           4   351    93    31     3    19    92
##           5    37     0   704     2  1408    60
##           6     0     0  1241     4  3068   162
```

In an attempt to remove noise from the data, we perform hierarchical clustering on the first two principal components, and compare them to the original clusters performed on all the data. The dendograms depict clear differences between the two. When all the data is used, the clusters end up relatively close to eachother, but when using the principal components they come from opposite ends of the plot. The table below also portrays the differences between the two clustering methods.

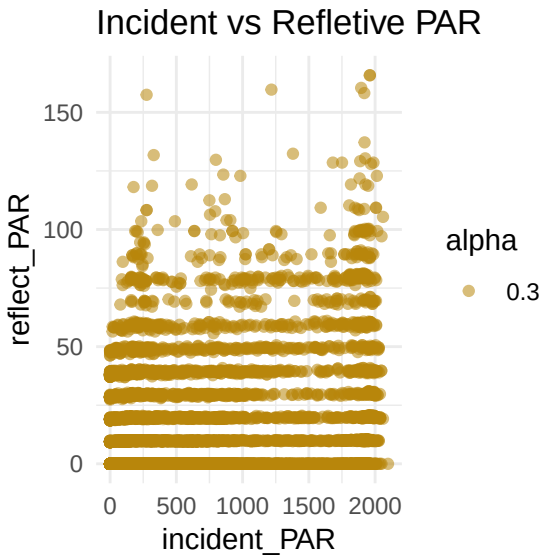



```
## hc.clusters
```

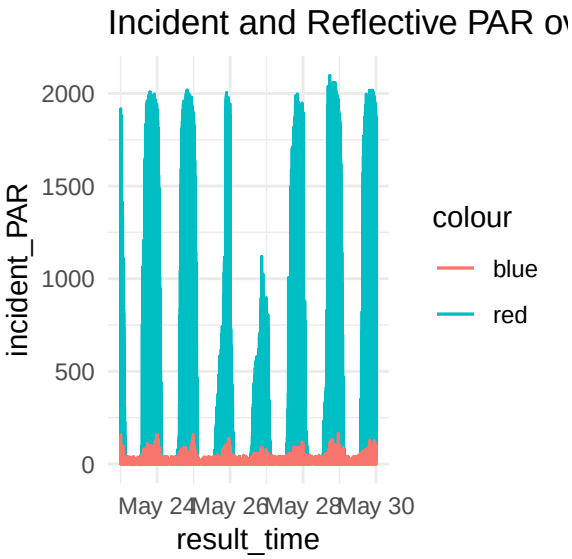
```
##      1      2      3      4      5      6
##   1179    604   6449     98  19124    472

##
##      1      2      3      4      5      6
##   3546   2287     68    601  20764    660
```

2. Finding 2:

Another interesting finding is how little Incident and Reflective PAR(Direct and Ambien) seem to be correlated when you look at their 2 dimensional scatter plot, given that both measure energy available for photosynthesis and tell us about drivers for the carbon balance in the forest.

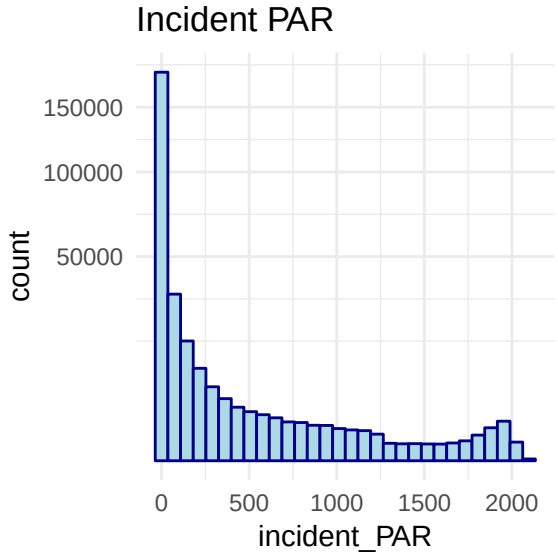


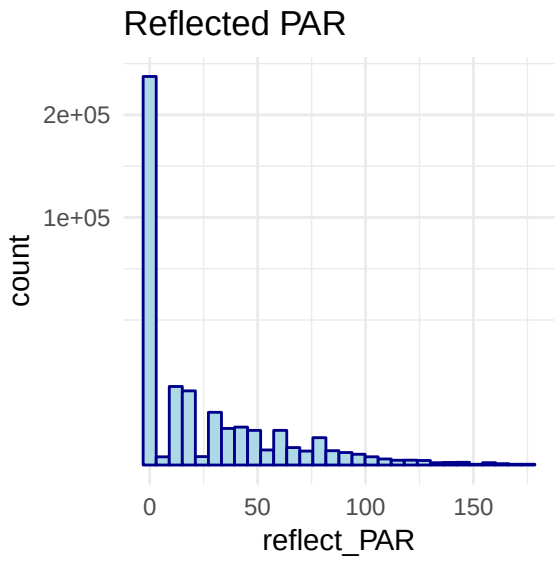
However, looking at both over time it's clear they experience similar cycles with the daylight, with the fluctuations of incident PAR having a much greater magnitude.



Graph Critique in the paper (40 pts)}
The overall quality of the paper by Tolle et al. is good. However, some plots are not perfect from a statistician's point of view.

- Figure 3[a] shows the distributions of sensor readings projected onto the value dimension, using a histogram. It turns out that both the incident and reflected PAR have long tail. We could not read full information from this histogram. Try to make a better plot with log transform of the data.





- What message do the boxplots in Figure 3[c] and 3[d] try to convey? Do you think the plots convey the right messages? If not generate a new plot with the same data. Hint: compare to some plots in Figure 4.

Figure 3c shows the distribution of the sensor readings taken on each of height, including median, quartile range, and outliers. Figure 3d shows the distributions of sensor reading differences from the mean. Both effectively convey what they're trying to. 3c conveys the relative consistency of the distributions for temperature and humidity, as well as the spatial trend of PAR over height. The figures show that there's outliers at the bottom of the canopy which occasionally receive full sunlight, while the higher parts generally absorb much of the light before it can reach the bottom.

*item Any suggestions for improving the first two plots in Figure 4? Can you distinguish all the colors in these two plots?

Although the figures are mostly effective in getting the point across, you cannot distinguish all of the colors in the first two plots of figure 4. One reason for this is the small size; it would be helpful to have larger figures and translucent lines.

*item Comment on Figure 7. Is it possible to generate a better visualization to highlight the difference between network and log data?

Figure 7 of the paper aims to visualize the data yield analysis of this study, by adding up the number of nodes at each timepoint that report a data value and dividing by the total number of nodes. While both **sonoma-data-log.csv** and **sonoma-data-net.csv** had issues during their deployment time periods, figure 7 shows us that both are important for collecting data. One alternative way to highlighting the difference between the network and log data is to also remove outliers from the data yield analysis. The multi-dimensional analysis technique used to study the delivery performance of the sensor network only removes missing values from the yield, but considers all reported values - including faulty ones. Removing these outliers could drastically improve this analysis, considering roughly ten percent of recorded values were faulty.

- 1: "10-bit Atmel Microcontroller with 128KBytes In-System Programmable Flash" http://ww1.microchip.com/downloads/en/DeviceDoc/doc2467.pdf
- 2: "Analog to Digital Conversion" https://learn.sparkfun.com/tutorials/analog-to-digital-conversion/relating-adc-value-to-voltage