Project 1 - Redwood Data Report

Yue Li Tianyu Wu

2022/10/13

1. Data Collection

(a) The paper studies the spatial and temporal dynamics of the microclimate surrounding a coastal redwood tree at different level heights. The paper considers temperature, relative humidity, and sunlight levels (measured by `incident PAR` and `reflected PAR`) as the primary factors for characterizing microclimate variations of the redwood tree. Data was collected for a period of 44 days on a 70-meter tall redwood tree. The team uses the "macroscopic" wireless sensor network by putting sensor nodes packaged in weather chambers on different sections of tree with variation in height, angular location and radial distance. This study improves the variation details of the above factors on temporal scale, since previous studies only showed that there are variations across spatial scale but not on temporal scale. The group collected data after the 44-day experiment period and analyzed the performance of the sensor network on capturing temporal trends of data to improve accuracy of future sensor deployment.

Here are some impacts we obtained from the paper on the sensor deployment:

The small variations in sensor positions (`height`) might lead to large differences in data collected, provided that the sensors are small enough and the phenomenon gets directional enough. As can be seen in the patterns of PAR data versus time, the data readings fluctuate according the sun's movement due to the belief that the wind moved the foliage and blocked sunlight access to the nodes. However, as can be seen in Figure 8, the patterns are actually consistent on different days. This contrast suggests that different orientations for each sensor result in different fluctuation patterns. The noisy data was a response from a highly-focused sensor.

To record the long-term performance of the sensor network for capturing data, network management is crucial. The paper suggests that one should include a network monitoring component that can evaluate the performance of the network timely and can report abnormalities. In the current study, the local logs ran out of space for data storage during the data collection period and led to network failure. Therefore, the network can serve as a way to detect and compensate for failures in logging. The logging can compensate for failures in the network.

The study explored the existence of spatial gradients in the microclimate and collected sufficient to see the variation of the gradients with respect to time. One can make use of the data for validating biological theories. Plant biologists can build a model of the effect of microclimate gradients on the sap flow rate to visualize the rate of sap flow varies over time with respect to humidity, temperature and PAR.

(b) Before deploying the sensors on the tree, the study performed roof and chamber calibration

checks. The calibration process evaluates performance data on different subsets of the sensors. Roof calibration establishes that PAR sensors are producing acceptable findings. Chamber calibration is a two-point calibration process to obtain responses of temperature and humidity.

The deployment package protects the electronics from the weather and safely exposes the sensors. All sensored are sampled every 5 minutes during one month in early summer, containing the most dynamic microclimatic variation. The nodes are placed on the west side of the tree. The nodes are placed 15 meters to 70 meters above ground level, with about 2-meter spacing between nodes, and at a radial distance of 0.1-1.0 meter from the trunk. Meanwhile, some nodes are placed out of the default angular and radial ranges to measure the microclimate nearby.

For data storage and management, the study included a local data logging system. The logger recorded readings taken by the queries before passing to multi-hop routing layer and stopped recording when the flash chip is full. The complete data logger was deployed since the capacity of the flash was sufficient for the duration of the data collection process.

The study first measures traditional climate variables: temperature, humidity and light levels. Temperature and humidity are parts of the transpiration model. The two variables connected sunlight levels are incident PAR and reflected PAR. Incident PAR quantifies the energy available for photosynthesis. In the dataset, these variable are: humidity (`humidity`), temperature (`humid temp`) and PAR measured on the top and at the bottom of the sensor (`hamatop` and `hamabot` respectively). `sonoma-data-log` dataset consists of data stored locally during the data collection process. `sonoma-data-net` constitutes data transmitted to the database during the data collection process. `mote-location-data` records the height, direction, radial distance and relative location to tree of the sensors. This dataset can help us analyze the temporal and spatial effect of the climate variables. `sonoma-dates` matches the epochs with the correct times.

2. Data Cleaning

(a) Variable voltage for two datasets are not in the same domain.

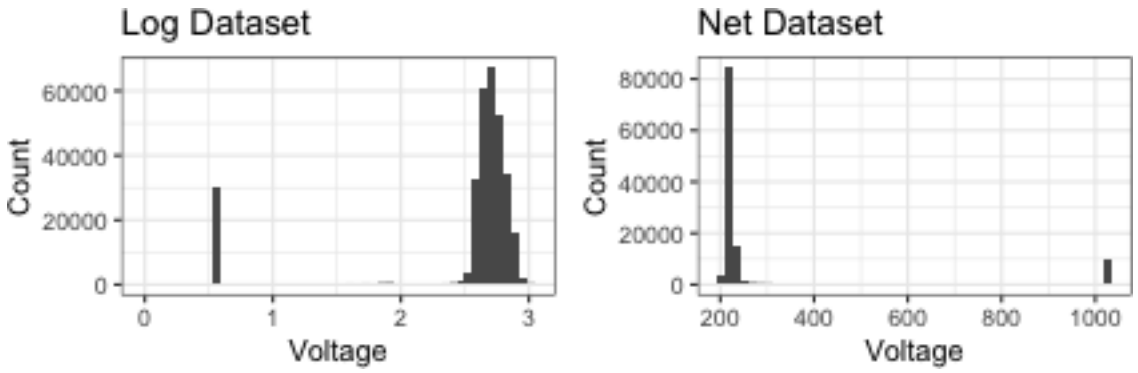

Figure 1: different scales for voltage

We know from the paper that the incident and reflected PAR measurements were collected by two Hamamatsu S1087 photodiodes interfaced to the 10-bit ADC on Mica2Dot. From user manual (http://www-db.ics.uci.edu/pages/research/quasar/MPR-MIB%20Series%20User%20Manual%207430-0021-06_A.pdf Page 25) we found the formula of conversion to be $V_{batt} = V_{ref} \times$

ADC_FS/ADC_Count. The voltage data from net dataset is converted to the right range after the conversion.

(b) It turns out that when a row contains any missing values, all four variables of interests are missing. Missing values are found in between 2004-04-30 and 2004-05-25 (27 days) for `log-dataset` and in between 2004-05-07 and 2004-05-29 (22 days) for `net dataset`. These two time period almost span the entire experiment. Most of the time there is only one record with missing value for each epoch, which indicates that most of the time only one single node is not working well. The bar plot of count of missing values for each epoch is not attached to save space. In the end, 8270 rows of missing values are dropped from `log dataset` and 4262 rows are dropped from `net dataset`.

(c) First, we combine log data and net data together. The general strategy is outer join to keep as much data as possible. Given the fact that time span of log dataset covers that of net dataset and net dataset has more data records, we decide that when both dataset contain the same data record, we keep the one inside net dataset. The next step is the inner join with location dataset. After joining there are 327228/333319 rows remaining.

(d) First we extract some nodes with abnormal readings and verified the finding in the thesis that abnormal readings are correlated with abnormal voltage.

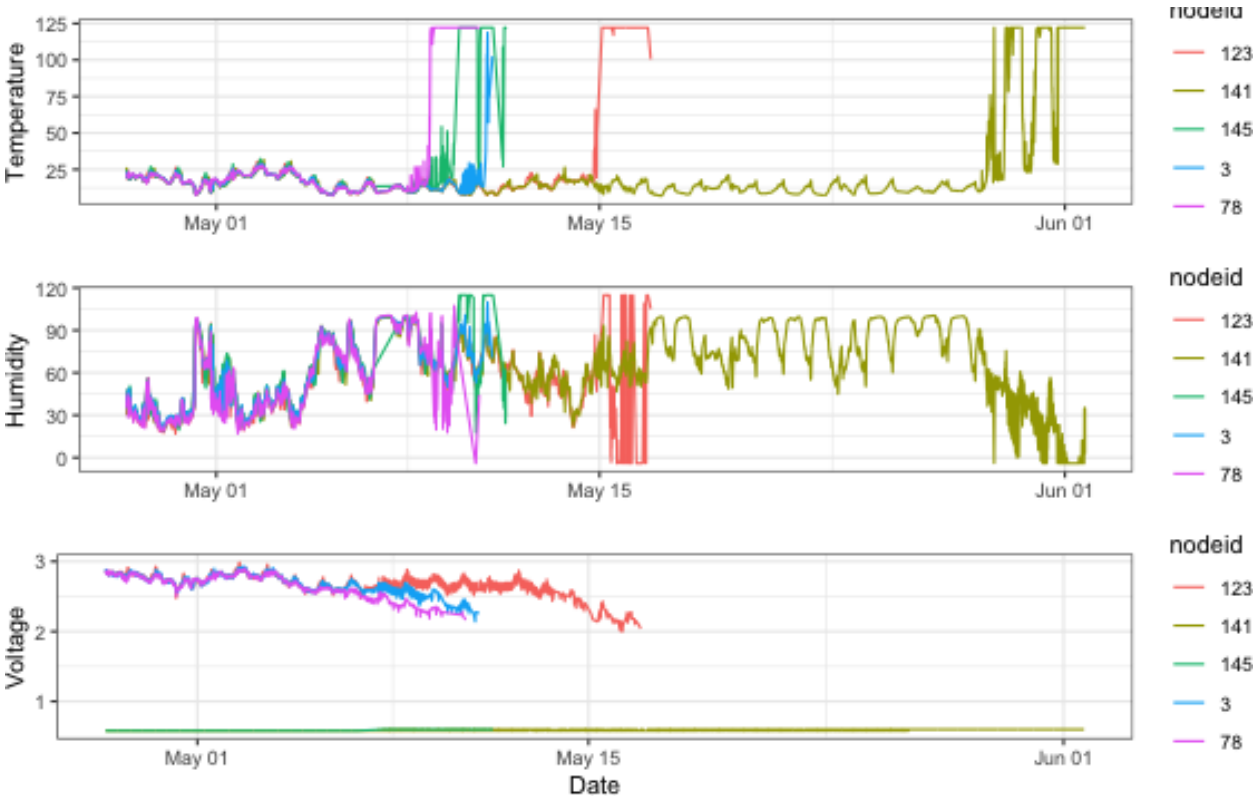

Figure 2: correlation between abnormal readings and low voltage

After removing observations with voltage lower than 2.4 and larger than 3, we manually identified some easy outliers, such as 2 temperature outliers from node 123, 3 humidity outliers from node 118 and incident PAR outliers from node 40. After checking all 98 readings from node 40, we decided to remove node 40 entirely as node 40 fails to operate almost at the beginning of the experiment.

There are obviously no more outliers for voltage, temperature and humidity from histogram. For rest of the two variables, even if we removed observations with reading 0s to make the plot look better, the histograms are still not convincing enough. To check the behavior of the long-tailed incident and reflected PAR, we plot the overall trend of PAR against time for some of the nodes with extremely high readings. All nodes are perfectly aligned with almost the same absolute value, which indicates that the readings are valid. Some histograms are not attahced to save space.

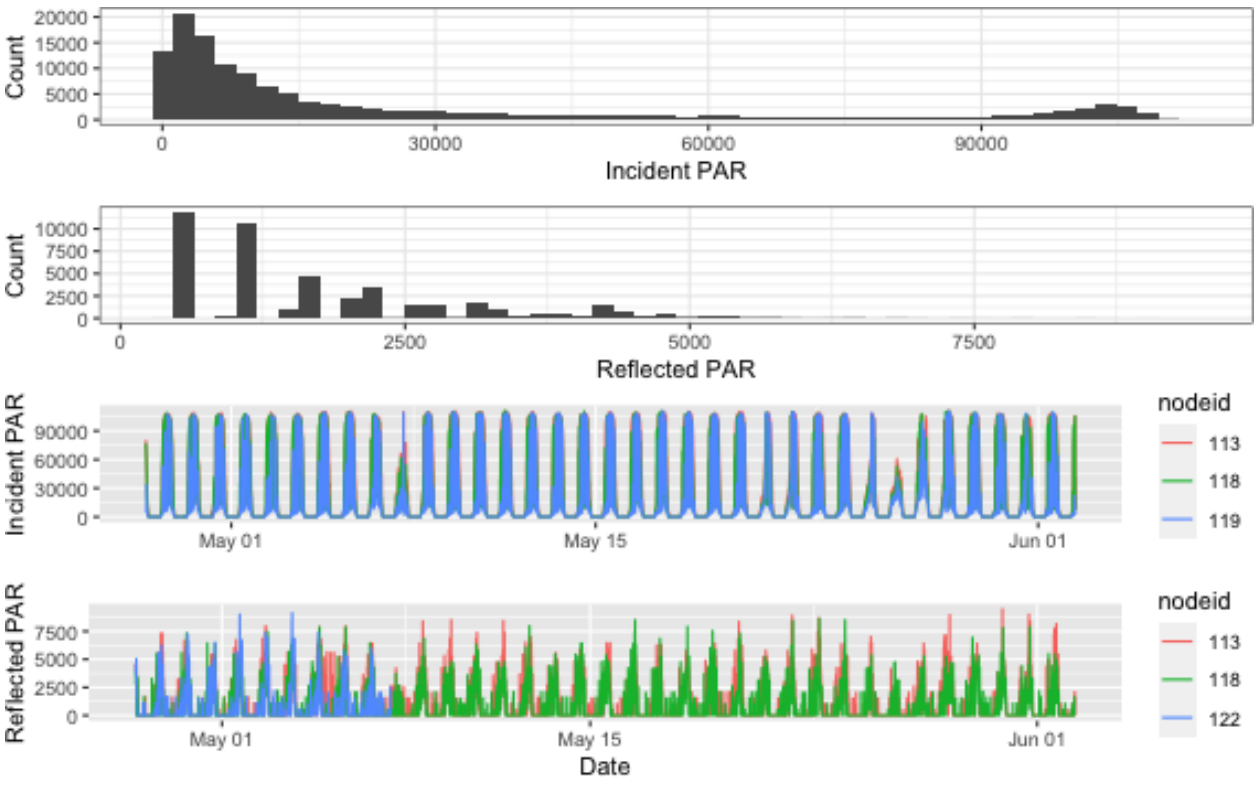

Figure 3: histogram after outlier objection and large-reading nodes' pattern against time

(e) Besides the outliers removed above, there are other outliers as well. From the analysis above we see that nodes tend to generate abnormal readings when their voltage runs low right before they are dead. When the reading is not extremely absurd, they cannot be detected by histograms/summary statistics but they are indeed problematic. We identify and remove some of this type of outliers by comparing the deviation of each node from the average reading of each day. We only consider the case that large deviation is at the last day certain node generates readings.

3. Data Exploration

(a) We decide to present two scatter plots that show significant linear relationships: `Reflected PAR` versus `Incident PAR`, `Humidity` versus `Temperature`. The time period chosen is from beginning to 2004-05-26. The reason for this choice of time period is that significant number of nodes become inactive after 2004-05-26 (elbow point) and that the relationship remains similar for future time periods. The plot shows a positive relationship between reflected PAR

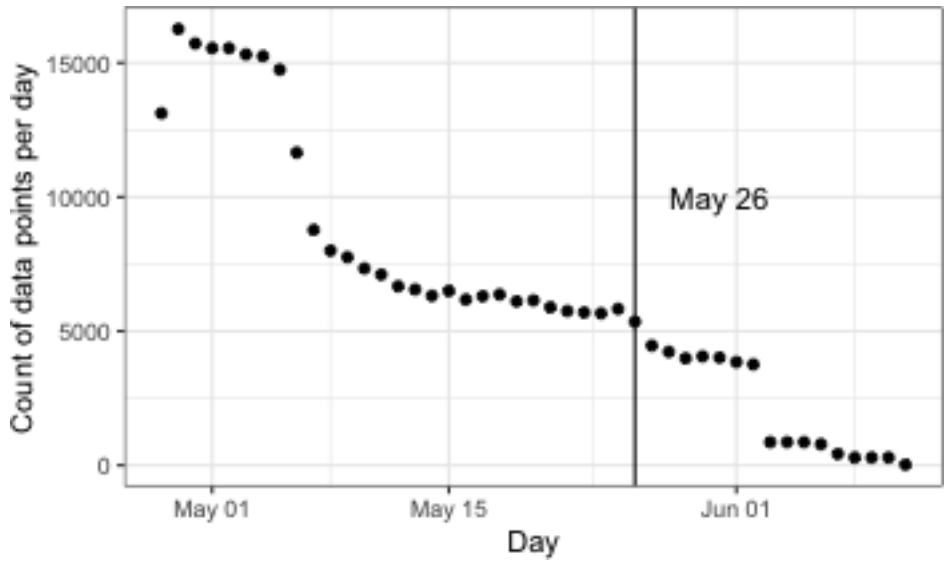

Figure 4: period selection

and incident PAR. This follows our belief, since both variables are measuring the amount of sunlight. However, as the scatter points indicate, the true relationship between these two variables may not be linear. The plot shows a strong negative relationship between relative humidity and temperature. This is consistent with the observation mentioned in the paper: warm days are dryer and cold days are more humid in the experiment site.

(b) There seems to be a positive linear relationship between Incident PAR and height. This is consistent with our common sense that higher nodes receive more sunlight. Lower height nodes have significantly fewer high values for incident PAR. Moreover, reflected PAR and incident PAR are positively correlated from the analysis in last part.

(c) We use time series analysis with different height levels to reflect the general trend for temperature, humidity, reflected PAR and incident PAR. First, we use daily mean to summarize the data. From the figure, we see that the higher the nodes, the more they are related to higher temperatures, incident and reflected PARs. Lower nodes tend to have slightly higher relative humidity. Second, we use hourly mean to summarize the data. Conclusions are similar regarding height. Temperature reaches peak around noon and the peak of PARs turn out to be around 3pm in the afternoon. Plots for hourly trend are omitted to save space.

(d) Four variables are selected for PCA: `humidity`, `temperature`, `reflected PAR`, `incident PAR`. From the screeplot, this data can be approximated by low-dimensional representation. According to the rule of elbow and Kaiser's rule, we see that the first two PCs explain 84.6% of the total variance, so they serve as a good low-dimensional representation.

4. Interesting findings

(a) We apply Gaussian mixture model to environmental variables: `humidity` and `temperature`. We see that humidity and temperature trends are very similar within one day in May. For the sake of convenience, we choose a day in May (May 7th) for observation. The reason for

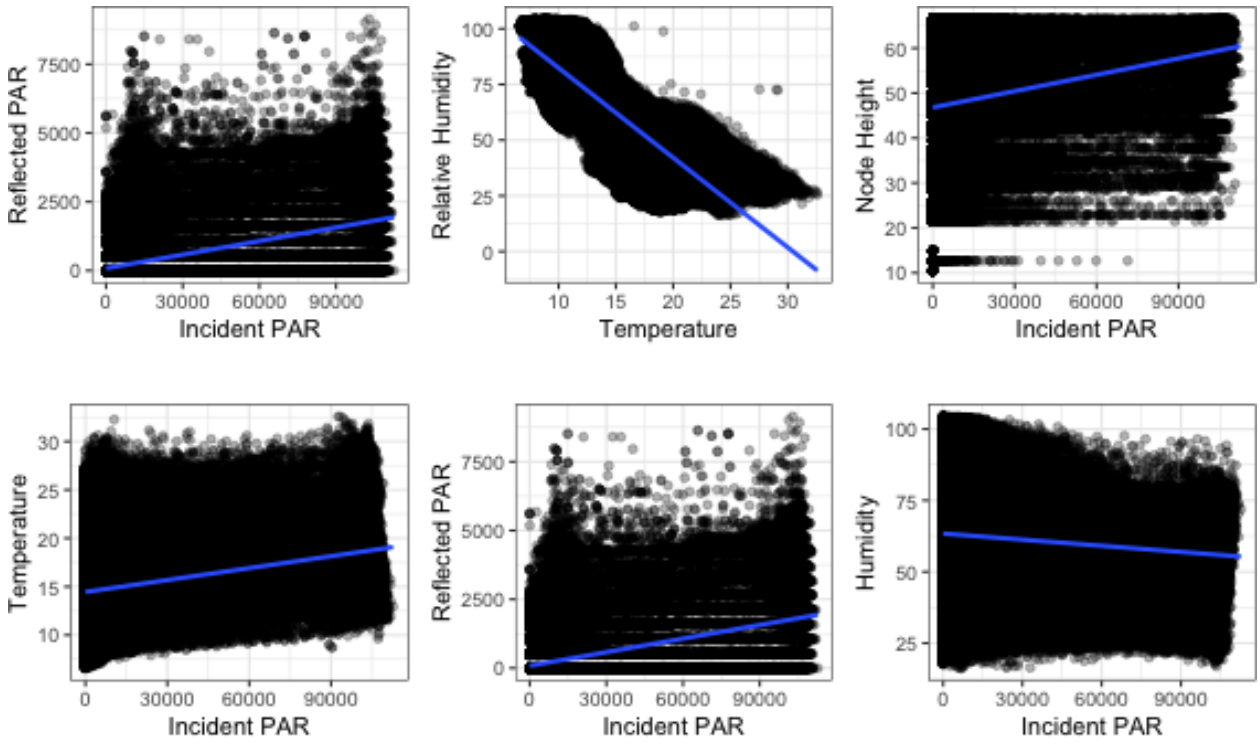

Figure 5: selected scatter plots

picking this date is we have shown that the trends for temperature versus humidity are similar across different days between May 6th and May 8th. The scatter plot is omitted to save space. From the density curve plot, Gaussian mixture model divides the observations into two groups: one with high temperature and low humidity, the other with low temperature and high humidity. To figure out the reason for such division, we create a scatter plot on height versus time. There is no evident height difference between these two groups, but there is an obvious cutoff at 8am and 9pm, which correspond to the boundary between daytime and nightime. This shows a logical but indeed interesting finding: high temperature and low humidity before sunset, and low temperature and high humidity after sunset.

(b) We apply PCA to the four variables with all the observations after scaling. From 3(d), the first two PCs explain the majority of the information. From the PC scores, `humidity` and `temperature` are more important for PC1. `Incident PAR` and `reflected PAR` are more important for PC2. From Figure, we see `humidity` and `temperature` almost have the opposite effect on PC1 and PC2. This is consistent with the fact that `humidity` and `temperature` have a negative linear relation, as shown in Part 3.

(c) By K-Means clustering, `humidity` and `temperature` have inherent relations to the height of nodes. We compute the averages of `temperature` and `humidity` by height. Then we start clustering with two centroids and then look at the distribution of height by cluster. The distributions show differences in height, so there are inherent relations in between. The same reasoning applies for `reflected PAR` and `incident PAR`, as well as these pairs against the time of recording.

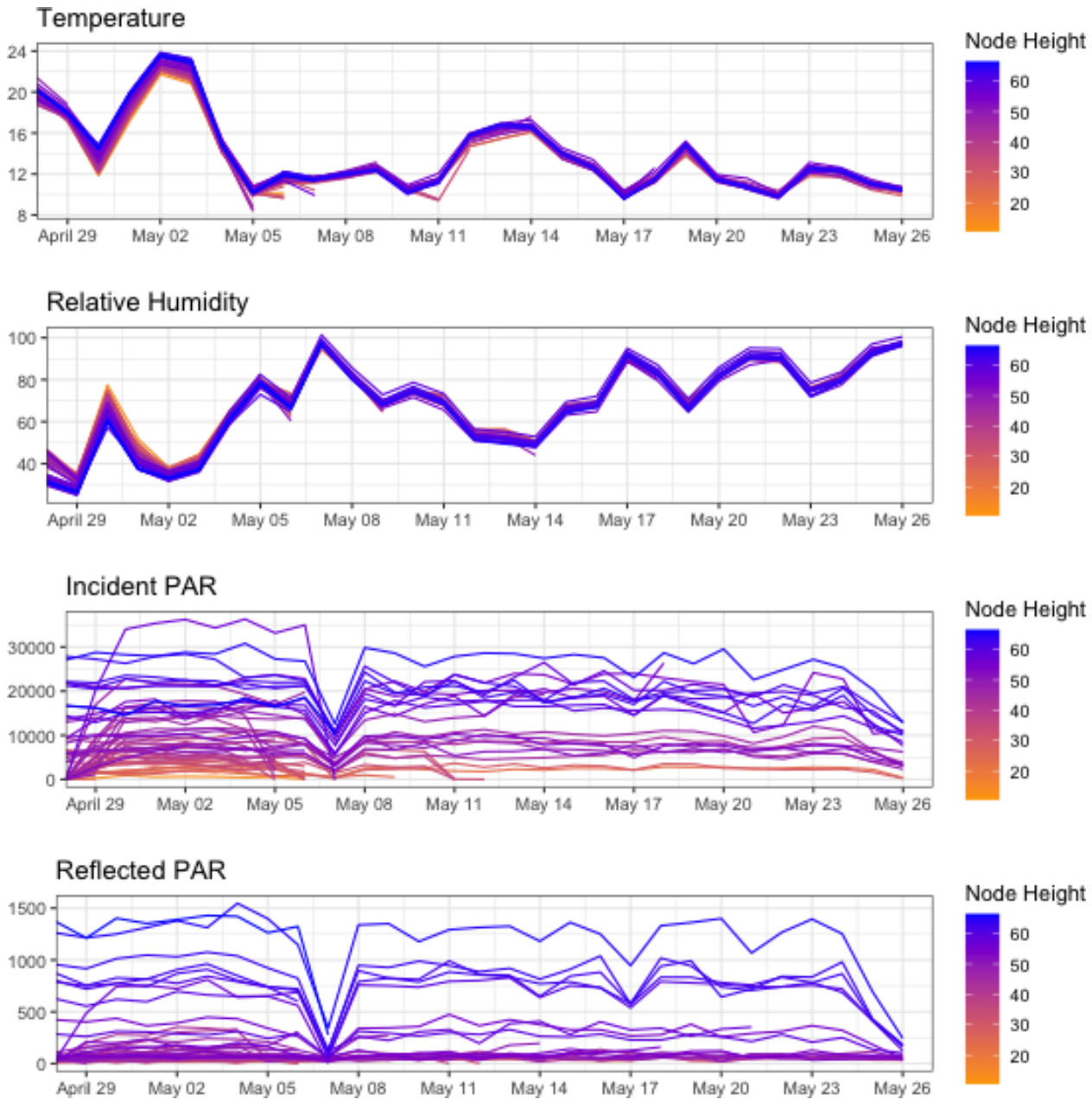

Figure 6: time series plots for the four environmental variables

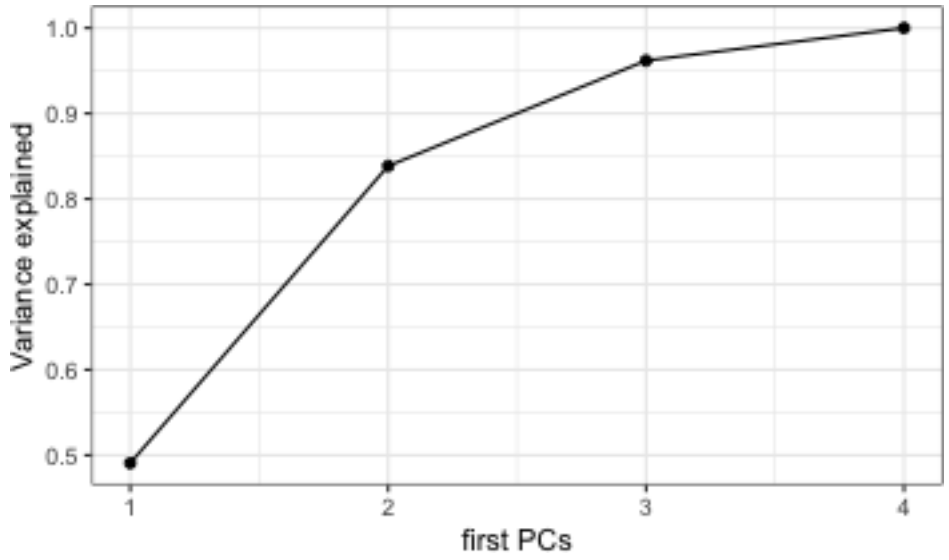

Figure 7: screeplot

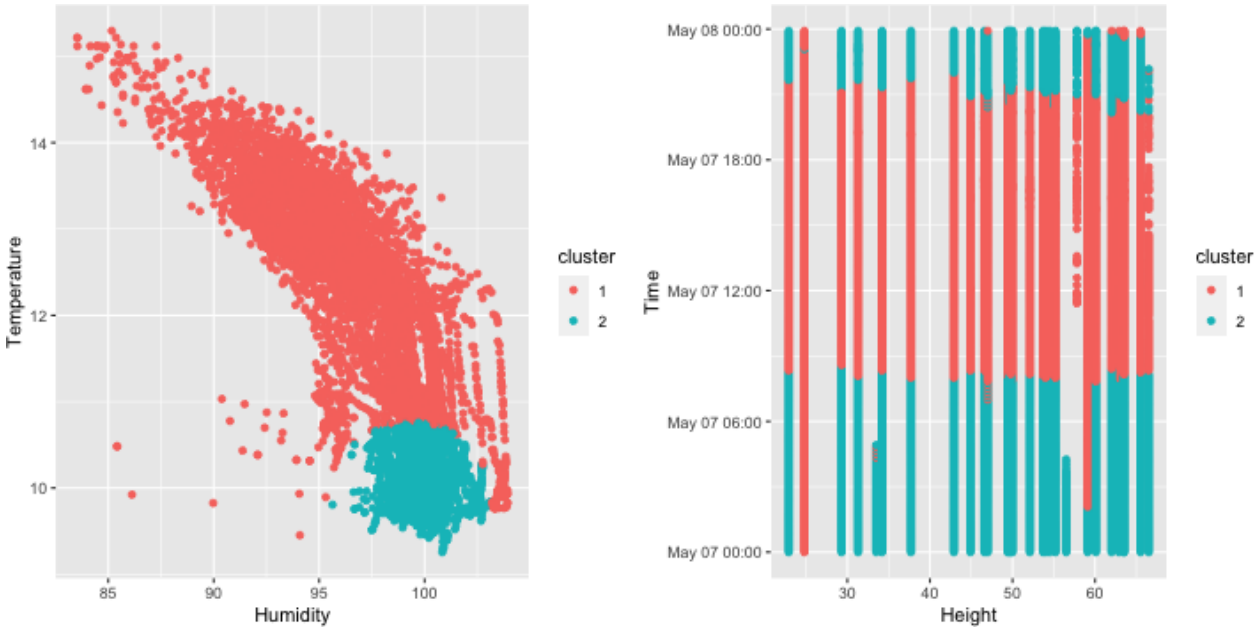

Figure 8: gmm grouping result

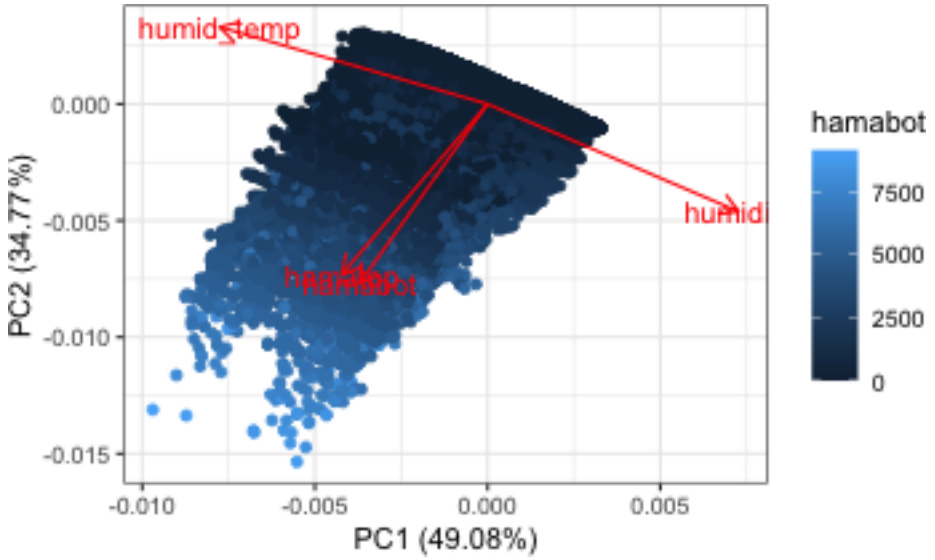

Figure 9: pca

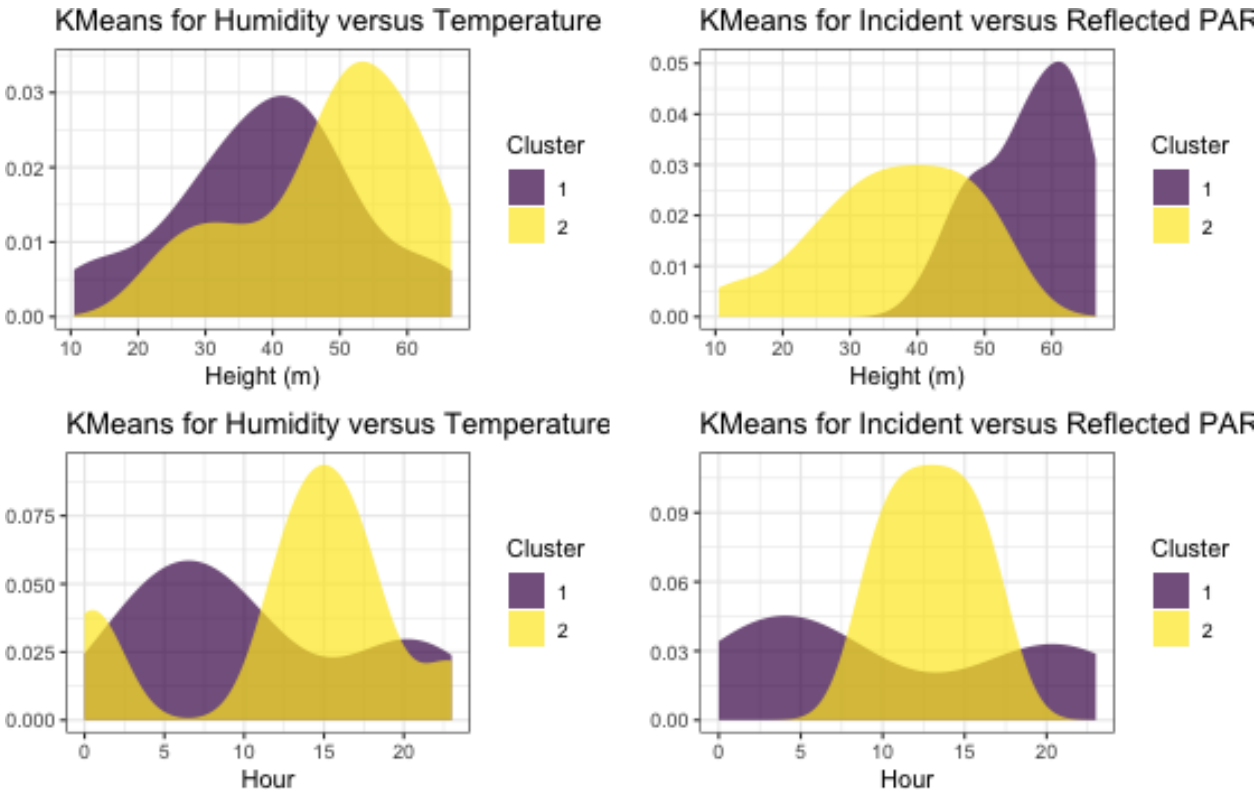

Figure 10: k-means against height and time

5. Graph Critique

(a) Since there are too many zeros in `incident PAR` and `reflected PAR` that are not going to help with tail analysis, we filter all the zeros in the data and take log for transformation.

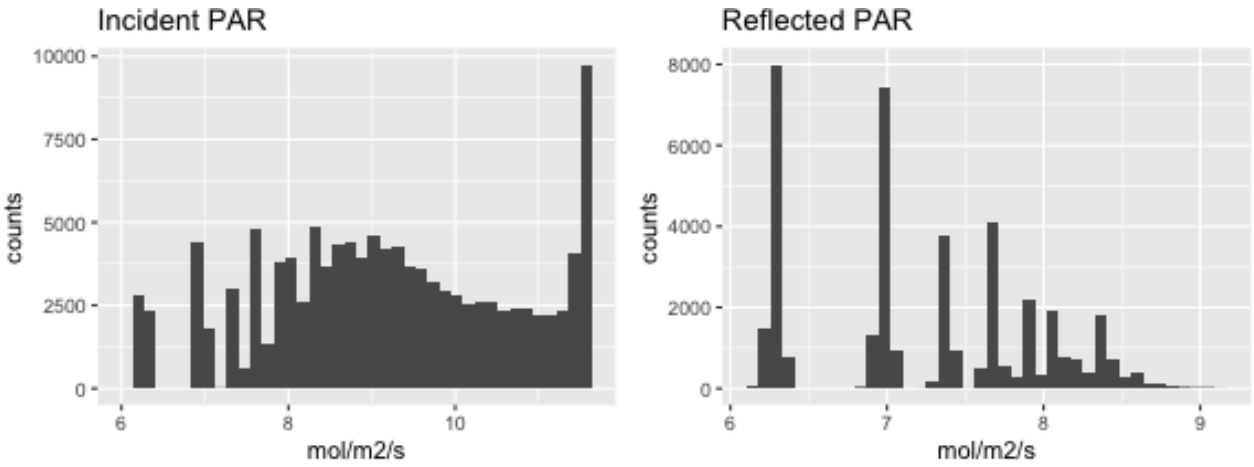

Figure 11: log transform of data

(b) From our understanding, the box plots in figure 3(c) and 3(d) in the paper show the distributions of different variables over height. The paper suggests that spatial gradients might exist over the height of the tree. However, we don't think the box plots convey a complete message without observing the temporal readings. Moreover, the difference brought in by temporal factors may cover the difference brought in by spatial gradients. Thus, we might as well simulate figure 4 presented in the paper to plot trends of the mean values of four environmental variables against height over different time periods during a day, and create several trend plots over different days. Each of color in the plots on the right represents different dates. Line plot turns out to be a better option to convey trends than box plot. We can see the spatial gradients clearly over time: `humidity` and `temperature` change across time periods within a day as well as the time periods across several days, but the trend over height is not obvious; `Incident PAR` and `reflected PAR` change across time periods within a day as well as the time periods across several days, and the trend over height is obvious.

(c) The disadvantages of the first two plots are that there are too many lines in the plot. A way to simplify the plot is to divide the nodes grouped by different height levels and plot the change of `humidity` over time. We can also simplify the plots for `temperature`, `incident PAR` and `reflected PAR`. The plots for PAR are not attached to save space, even though their trend against height is more obvious.

(d) We can concatenate the fourth plots of (a) and (b) respectively and combine the bars with the same height value. This will allow us to see the overlap between `net-data` and `log-data`. We can improve the third plots of (a) and (b) respectively by using bar plots instead of scatter plots. The second plots of (a) and (b) feature too many box plots which lack interpretability. A way to enhance interpretability is to partition the days into several groups and create boxplots based on the grouped result. The first plots of (a) and (b) respectively have the worst interpretation: the meaning of the x-axis is very vague to readers.

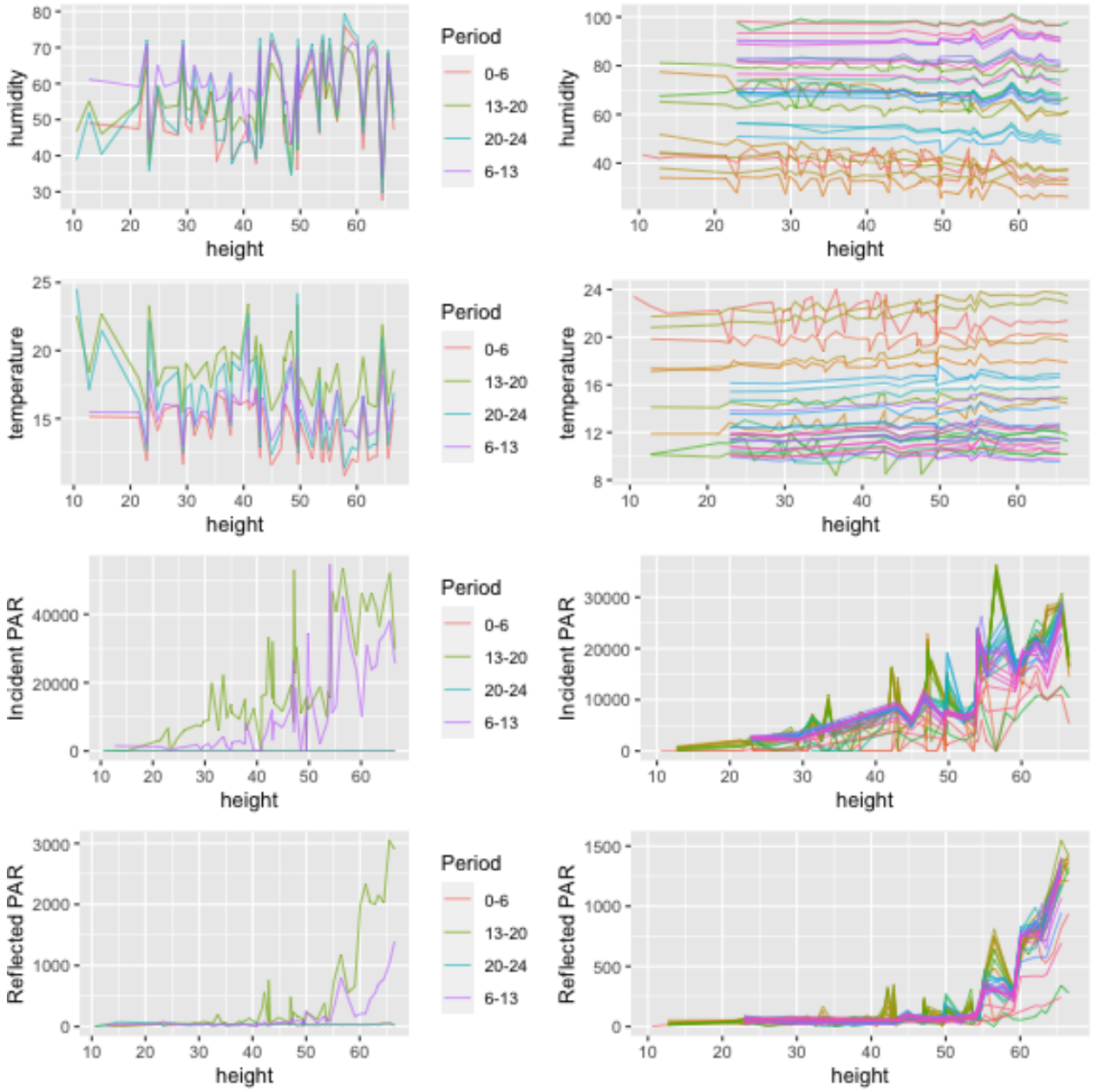

Figure 12: mean of four enviromental variables during a day and across different days

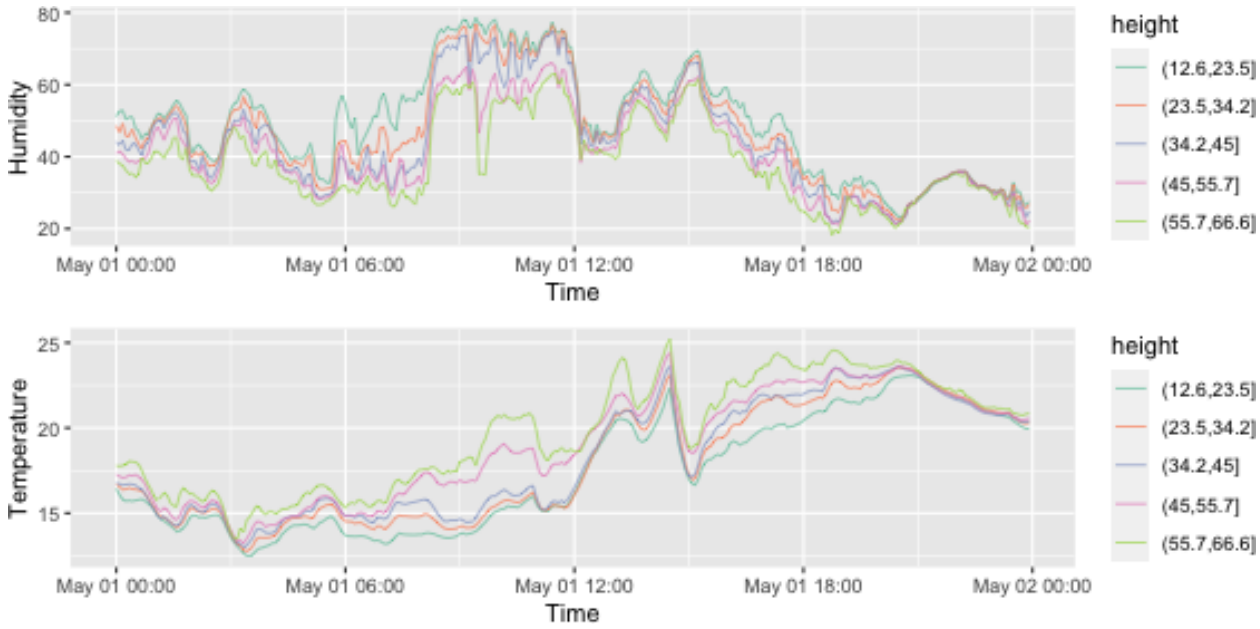

Figure 13: plot simplification

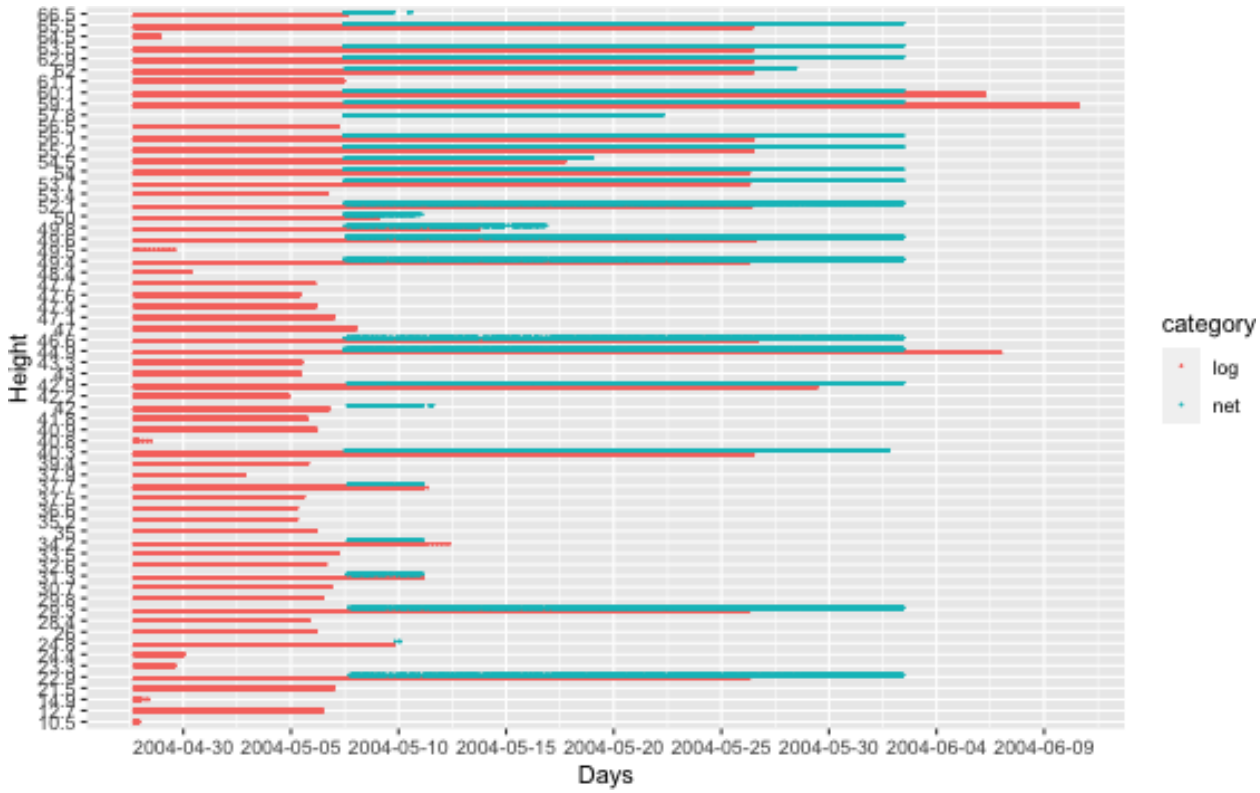

Figure 14: a better visualization for net and log data