STAT 215A
Lab 1: Redwood Data Lab
Timothy Meyers

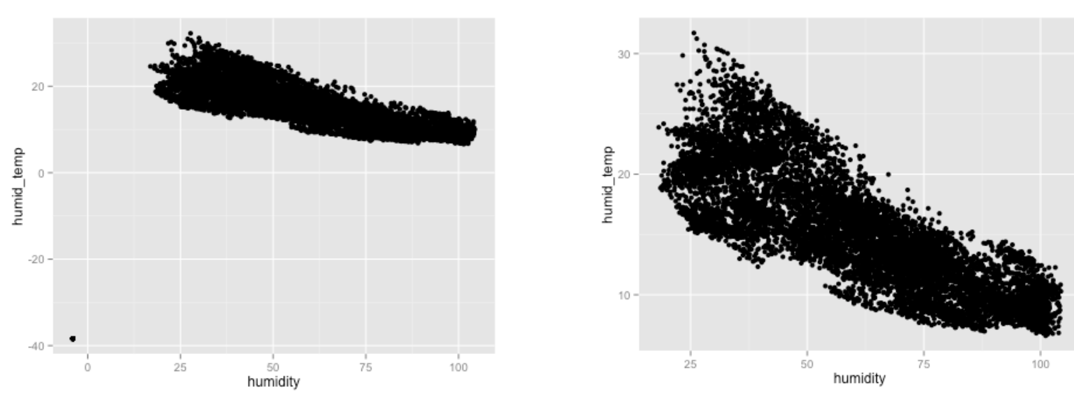
Exploration of Data

Variables

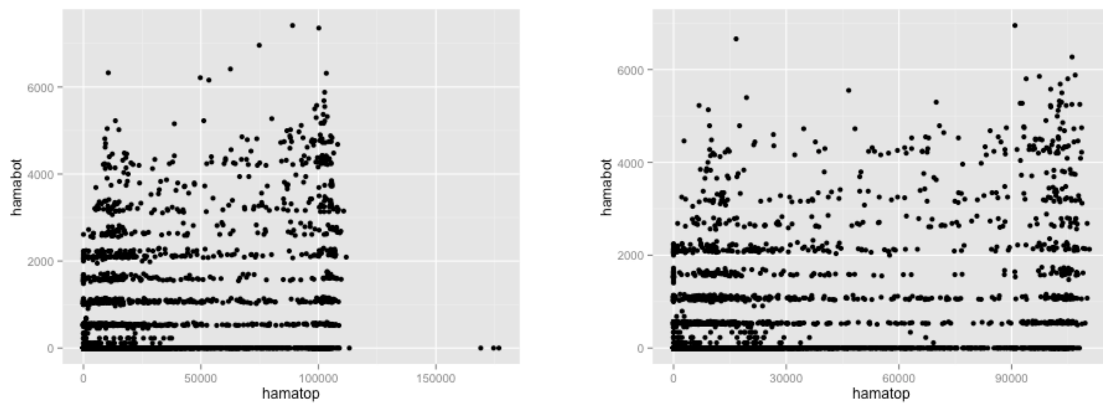
The data is measured from a series of sensors which collect temperature, humidity, and luminosity readings. The Sensirion SHT11 is a digital sensor that records both temperature and humidity. The measurement error for each of these variables is ± 0.5◦C and ± 3.5 percent, respectively. Humidity is greater than zero, but can be greater than 100 percent in foggy conditions. Temperature readings, even if they are within normal ranges, can deviate significantly from historical values depending on the sensor's battery. Voltage is recorded and provides an indication of the reliability of temperature recordings, and it's advised to focus only on temperature readings generated from a sensor with voltage between 2.4 and 3. Luminosity is recorded as Photosynthetically active radiation (PAR), and both incident (direct) and reflected (ambient) PAR are recorded. These measurements are collected by two Mamamatsu S1087 photodiodes, and are referred to as hamatop and hamabot in the data. The units of measurement are unclear.

Data Cleaning

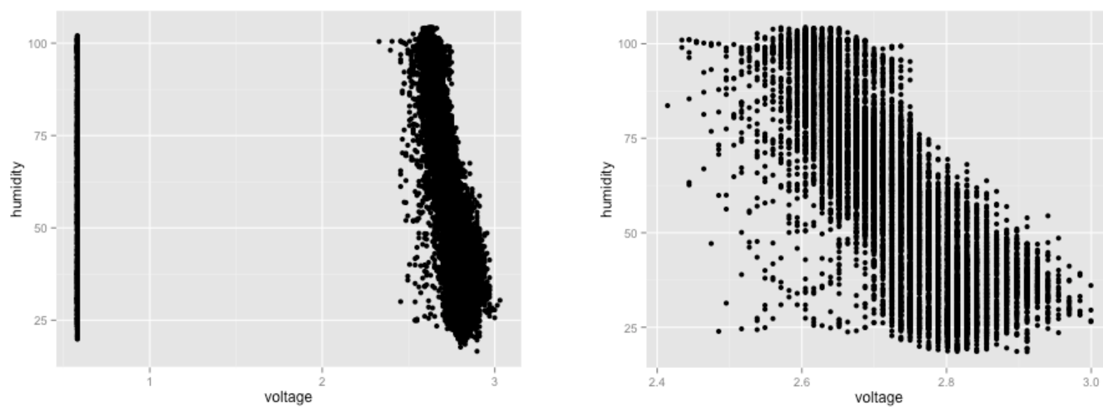
After loading the data and making a few plots, I saw that some rows had values missing. Also, looking at the plots helped me understand the ranges where values were expected to lie, allowing me to remove obvious outliers.



Removing outliers from humidity and temperature data

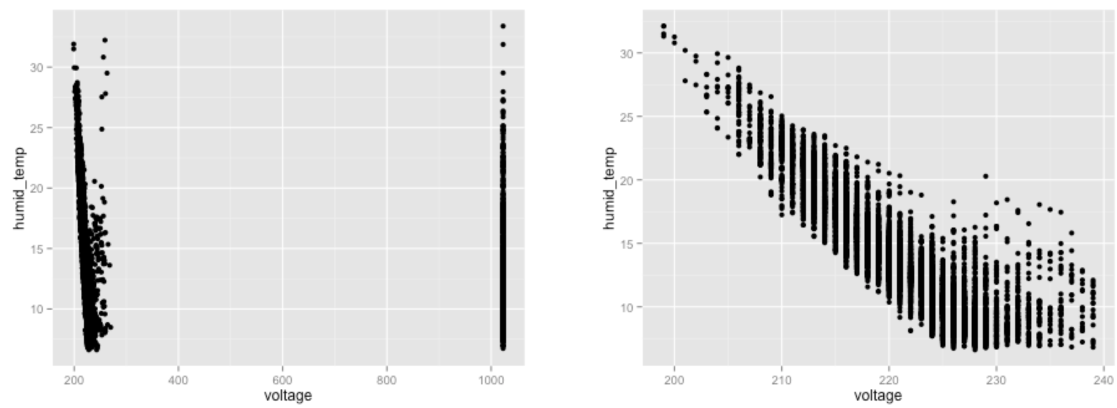


Removing outliers from reflected and incident PAR data

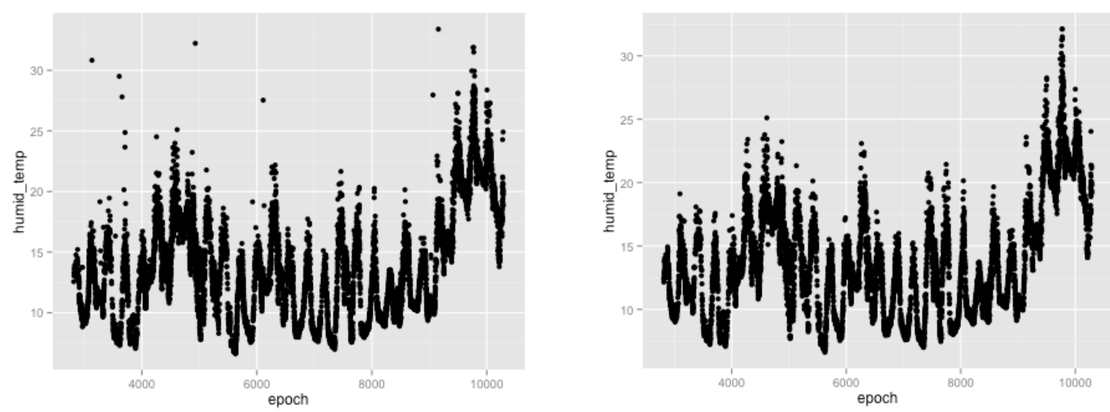


Removing outliers from voltage data

After applying basic data cleansing to the net and log data, I ran humidity and temperature versus epoch on the net data and found that there were quite a few outliers in the temperature measurements. These became more apparent when plotting against voltage, and I decided that 240 was the appropriate voltage cutoff to remove these outliers.



Removing voltage outliers from net data

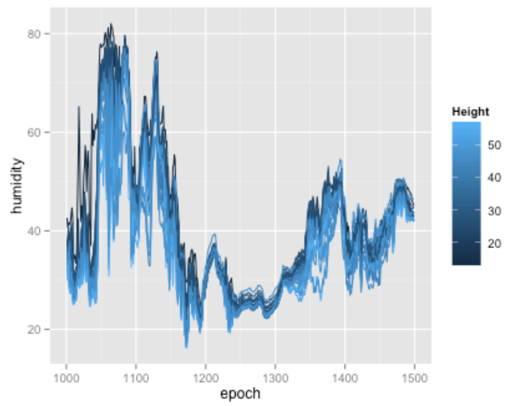


Temperature after removing voltage outliers

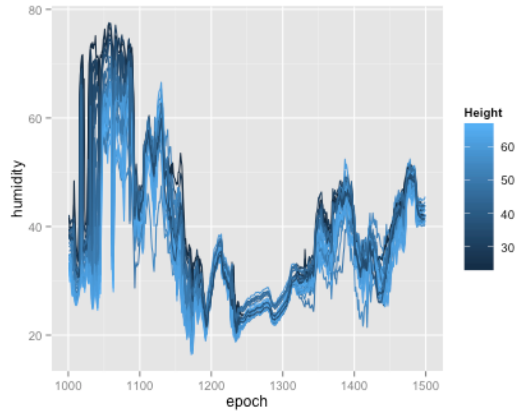
Let’s now take a look at location data, and how the position of the sensors impacts observed values.

Data Exploration

Let’s examine temperature variability for edge motes and interior motes across height.

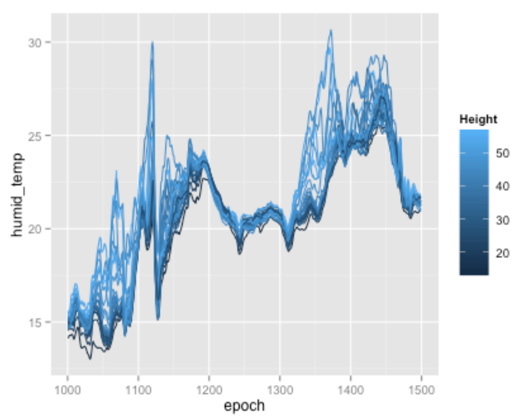


Humidity at tree edge across day

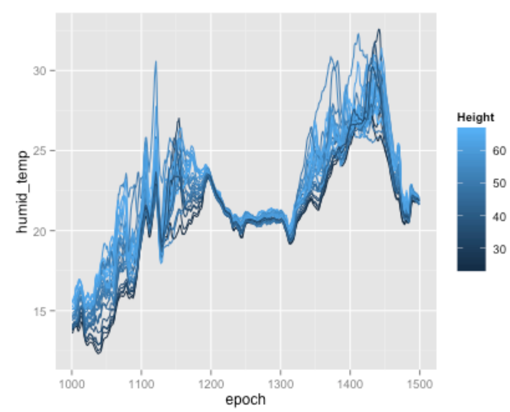


Humidity at tree interior across day

Above are humidity at the edge and interior of the tree over time, respectively, colored by height.



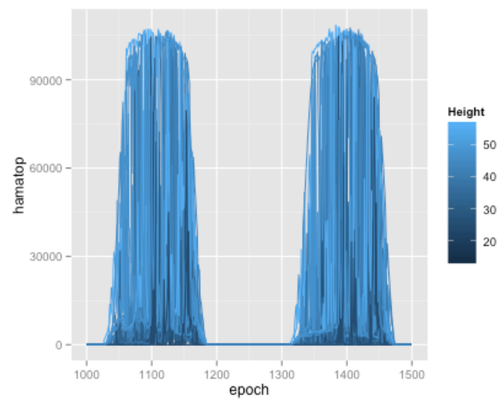
Temperature at tree edge across day



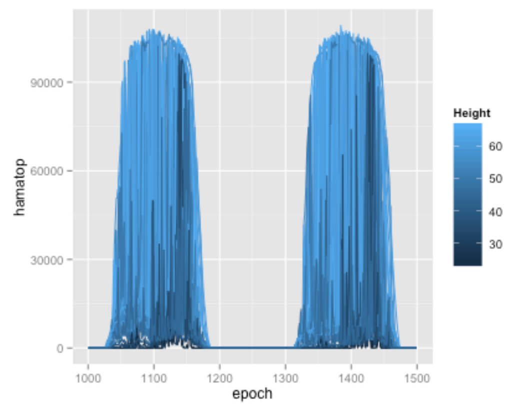
Temperature at tree interior across day

Above are temperature at the edge and interior of the tree over time, respectively, colored by height.

Two interesting observations are that temperature appears to increase with height, and that humidity appears to decrease with height. We'll explore these more later.

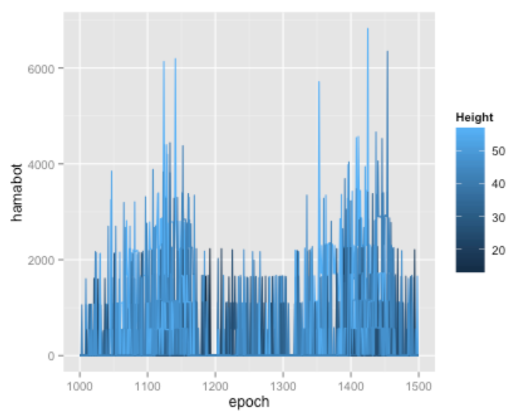


Incident PAR at edge of tree

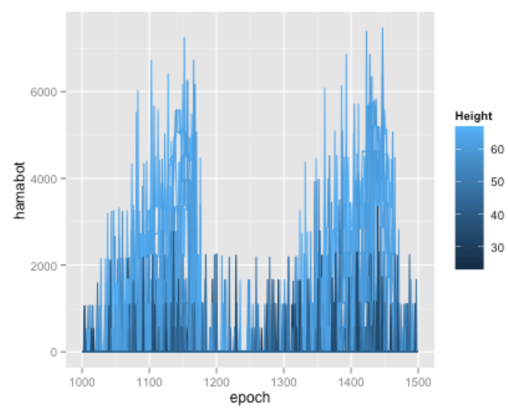


Incident PAR at interior of tree

Above are incident PAR at the edge and interior of the tree, respectively, colored by height.



Reflected PAR at edge of tree



Reflected PAR at interior of tree

Above are reflected PAR at the edge and interior of the tree, respectively, colored by height.

Now, let’s check out direction, regardless of edge or interior.

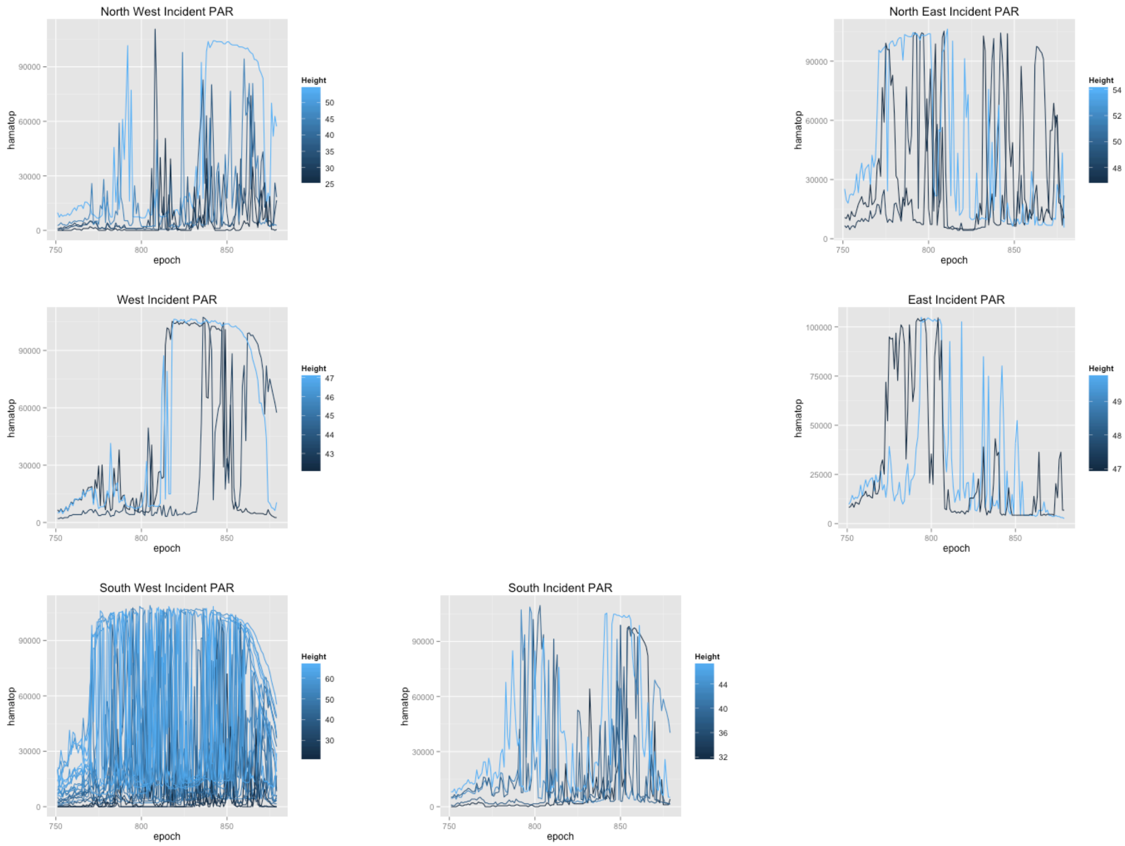


Figure 1 Incident PAR by Node Direction across Day

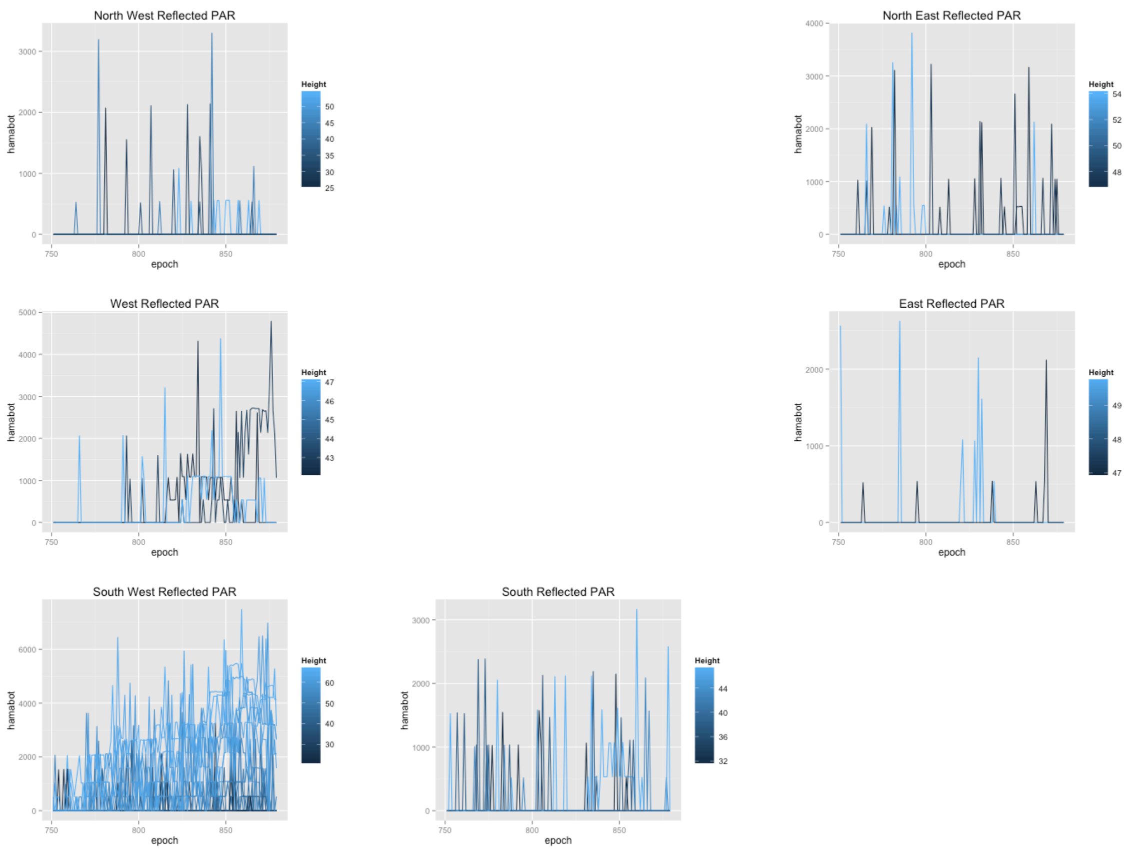


Figure 2 - Reflected PAR by Node Direction across Day

The above charts are positioned according to the position of the nodes they represent (i.e. southwest nodes are at the bottom left of the grid, east nodes are on the right of the grid, etc.) For the given day, southeast nodes are missing, and no node is positioned north.

Interpreting these charts, we observe that most nodes are located on the southwest side, and they receive the most sunlight.

Graphical Critique

The authors had two goals: 1) exploring "the spatial variation and temporal dynamics of the microclimate surrounding the coastal redwood tree", 2) analyzing "system performance to learn lessons for future deployments."

Figure 3 analyzes the data across several dimensions.

It first looks at the overall values observed. Figures (a) are straightforward, and do not pretend to do anything more than just understand the distribution of each variable. Figures (b) offer more insight, plotting the distributions from (a) along the time axis.

Figures (c) show the data cut across several other dimensions. In my opinion, they are slightly confusing, and not all useful. It took me a while to understand the y-axis; at first I thought it was interval data showing how variables changed across continuous height values. Afterwards, I realized they cherry-picked nodes with differing heights, so the y axis is actually nominal data with number labels. Also, the resulting graphs are difficult to interpret. Can we simply say that, besides for the node at height 64.5, there were no significant changes in weather? Perhaps differences only become significant during some interval during the day, so aggregating all temperatures across all times drowns the interesting results. Assuming the charts tried to ask: "Do temperature and humidity change with height?", I would say they do not successfully answer the question. The two other figures in (c) are better because they show a clearer relationship between PAR (incident and reflected) and height, despite the same flaws with the y-axis.

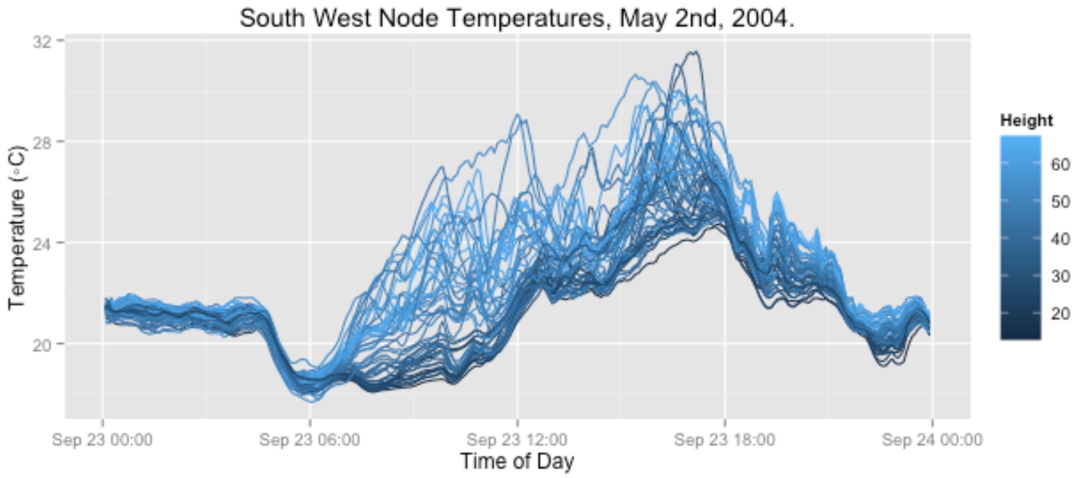
Figures (d) demonstrate where the variability in the data comes from. For temperature and humidity, it highlights the same problem that I mentioned as in their corresponding figures in (c). However, the charts do a better job at answering the question: "Do temperature and humidity change with height?", since we now see differences in temperature between top and bottom in the outliers. The PAR figures in (d) don't necessarily shed more light then their corresponding figures in (c).

Figure 4 takes a closer look at the variables, examining intra-day changes in the values. These figures are clear and self-explanatory, but could use slight improvements. The two top-left figures unnecessarily color every node – since we're more interested in the value spread, keeping all lines the same color would accomplish the same effect.

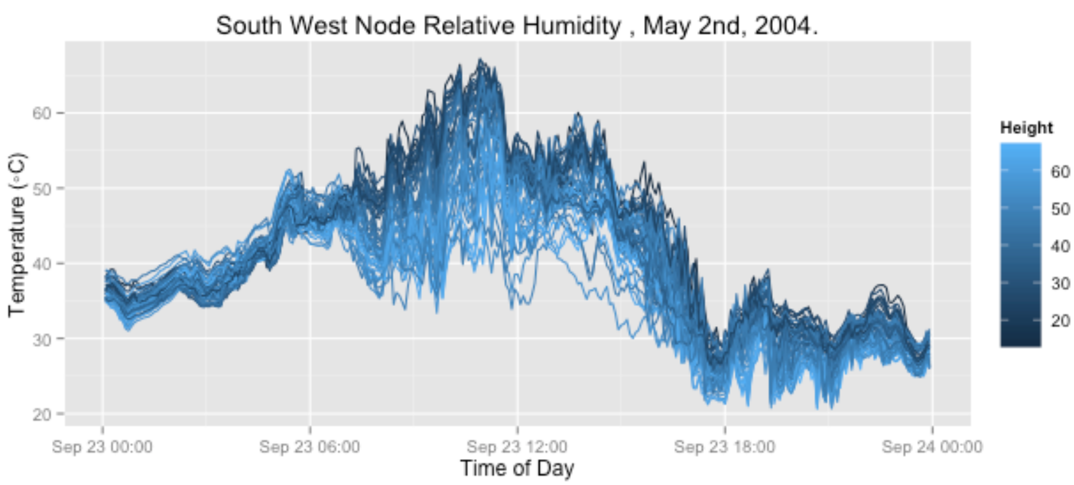
In my own analysis we saw that most nodes are on the southwest side, and these nodes receive the most sunlight. The report's authors remark that differences in temperature are due mostly to solar influence, which leads us to believe that node direction is a very important determinant for temperature recordings. It would have been interesting to layer direction into the figure instead of focusing only on height.

The PAR figures are interesting, but could also benefit from layering direction. It is unclear to me which nodes are included in the two bottom right figures. There are over 80 nodes in total, and we are only seeing a subset of these which may or may not be a balanced representation of direction, and hence of their ability to receive sun throughout the entire day.

Presenting Findings



The above figure plots temperatures recorded by motes positioned in the South West direction throughout the day. These motes receive the most sunlight. By ignoring other directions we can focus on height as the only factor in receiving less or more solar influence. We can see clearly that temperature increases with height.



The above figure plots relative humidity recorded by motes positioned in the South West direction throughout the day. We can see that humidity decreases with height.

Discussion

The data size definitely restricted me. When plotting data directly off a data frame, R Studio crashed. The workaround was to randomly sample and display a subset of the rows.