# Lab 1 - Redwood Data
# Stat 215A, Fall 2014

John Haberstroh

September 23, 2014

## 1 Introduction

Data from a macroscope for coastal redwood trees is considered. This data is to be cleaned, and trends are explored.

## 2 The Data

For our dataset, there are a number of static and dynamical variables. The static variables describe the identity of our sensor. This is comprised of:

- nodeid (appx 1-200)
- height (meters)
- distance (meters)
- direction (cardinal direction)
- starting time (date and time)

Dynamical varibales were measured at 5-minute intervals. These variables were comprised of:

- epoch, counting quantity of 5-minutes from starting time
- temperature (Celsius)
- humidity (percentage)
- indicent Photosynthetically Active Radiation (PAR) (nmol/m$^2$/s)
- ambient Photosynthetically Active Radiation (PAR) (nmol/m$^2$/s)

There were two datasets, log and net, collected locally and over a wireless transimission respectively.

## 2.1 Data Collection

For measuring ambient and incident light, the authors use Hamamatsu S1087 photodiodes with 10-bit depth and sensitivity to Photosynthetically Active Radiation (PAR). The device was constructued to have a lower compartment that shields from direct sunlight and an upper compartment that receives direct sunlight. The detector error was not specified, but it is likely that the error of nodes relative to eachother will be larger than a node relative to itself. Furthermore, hhe intensity of light at a particular position may vary faster than the measurement rate of one measurement/5 minutes; thus, these measurements should be examined and processed carefully to find robust indicators.

The humidity was measured with the Sensirion SHT11 digital sensor. This device provides precision on humidity up to ±3.5%. Humidity measurements should be more intrinsically robust than incident light, as the wind speeds inside of trees is not tremendous, and humidity will vary fastest with the aid of air currents.

## 2.2 Data Cleaning

To clean the dataset, I first merged nodes with location data, keeping only the nodes that had a specified height and tree. I only worked with the log dataset, because that dataset was larger after the merge by about a factor of two. I could not merge the two datasets directly because their dates were offset by several months. These sets could be combined if the seasonal effects are accounted for.

After this initial merge, I converted epoch and timestamp information into time.dmin, time.day, and time.year. This allows for a more familiar temporal analysis, and easy ways to slice data by day/time. For the computation, each epoch value was multiplied by 5 mins and added to time, and the resulting timeseries seem consistent with one another. Whether a systematic offset of 5mins is present in any sample is not clear, since weather patterns may have delays of a similar order of magnitude due to spaital variation within the tree.

To further clean the data set, I performed a range analysis to find quantities that were not valid. The invalid quantities discovered were (a) one node which produced only negative humidities for its entire lifespan and (b) one node which produced NA starting partway through its lifespan until its final entry. All nodes with negative humidities were entirely removed from the dataset, and all data was removed from days with NA entries on a day-by-day basis. There were also some humidities above 100, but these came from seemingly functional detectors, and were thus left in the dataset under the assumption that they came from calibration errors, and were just randomly distributed around the mean. No effort was made to modify the calibration.

For Incident PAR, outlier data was found near the beginning of the dataset. There, a much greater range of values was measured on all detectors. Examining this data closer, it seemed to have a character dissimilar from all other days.

This may be due to calibration from turning on. Because this first day started near noon and its behavior was uncharacteristic, these early data were removed.

### 2.3 Data Exploration

The distributions for incident light were quite complicated to interpret. The variability of the incident light was more exponentially distributed than gaussian distributed, so means had noise that was dominated by fluctuations to the tail. In photosynthesis, there are two main aspects of the response to light: first, the response of power generated to light is approximately linear until it saturates, and second, an excess of light causes photodamage. Thus, the most physically relevant analyses would consider separately the amount of power received by the plant and the amount of time spent in dangerous light. This type of analysis would separate the tail from the body of the distribution, and would likely lead to a robust analysis.

## 3 Graphical Critique

Next we will examine the plots presented from Figure 3 and 4. Figure 3(a) attempts to get a sense of the scale of data that we encounter, without considering time or height. This can verify that we are within the valid working range for our sensors. We can also get a general sense of the distribution of data: do the fluctuations within certain control parameters (e.g. time or height) dominate the local mean, or is the signal already demonstrating features relating to the local mean?

**Figure 3(a)** This plot has problems by representing qualitatively different distributions with the same method. This bar graph does very little to capture the distributions seen with only 10 bars. For both temperature and relative humidity, a kernel smoothing function would be more appropriate due to the fairly dense and uniform points. For incident and reflected PAR, the tremendous number of 0 values make the plot difficult to interpret. A log plot would be clearer, with all zeros binned together.

**Figure 3(b) and (c)** These plots deconvolute the effect of date and height, respectively. Averages are performed over the course of the day or over all days, respectively. Plot 3(b) demonstrates the importance of date for both temperature and humidity, while the importance of date for PAR is less clear. Plot 3(c) demonstrates the slight trend in temperature and humidity with height, and discovers a potential outlier/malfunctioning node at 64.5m. The significance of this trend is unclear, especially with the high variability in each measurement; the standard error of the mean could clarify the certainty of this trend. There is also a trend of increasing variability with height in both types of PAR. Again,

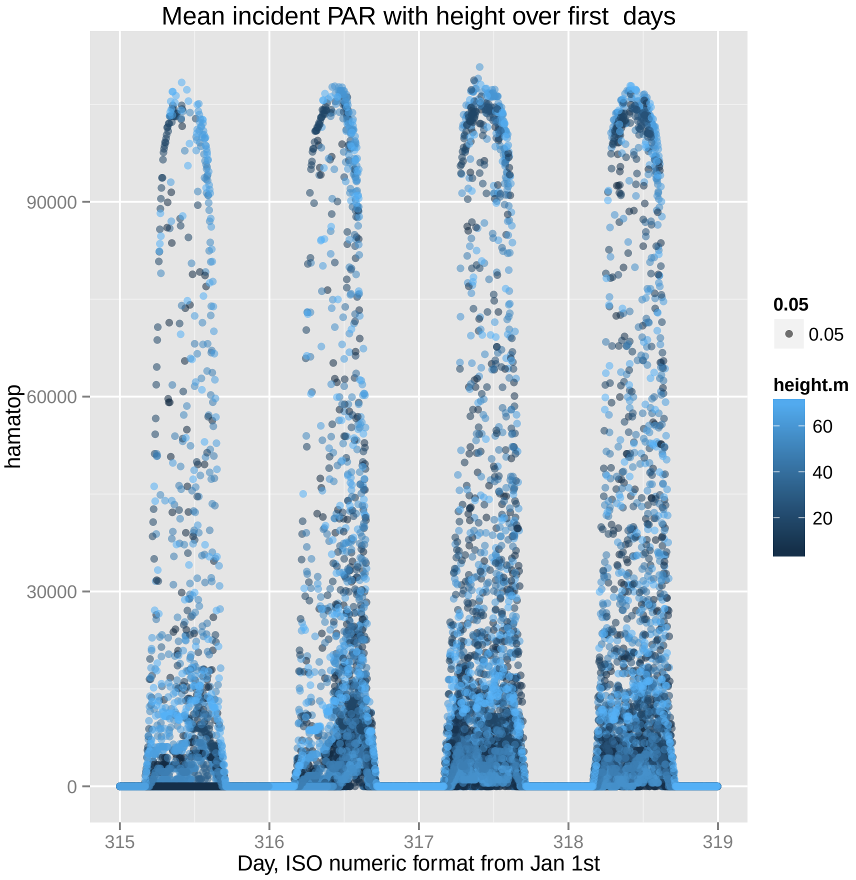

Figure 1: Incident light in time, with heights indicated by color. The higher heights are above the lower heights in many cases, but a simple averaging fails to reveal this trend.

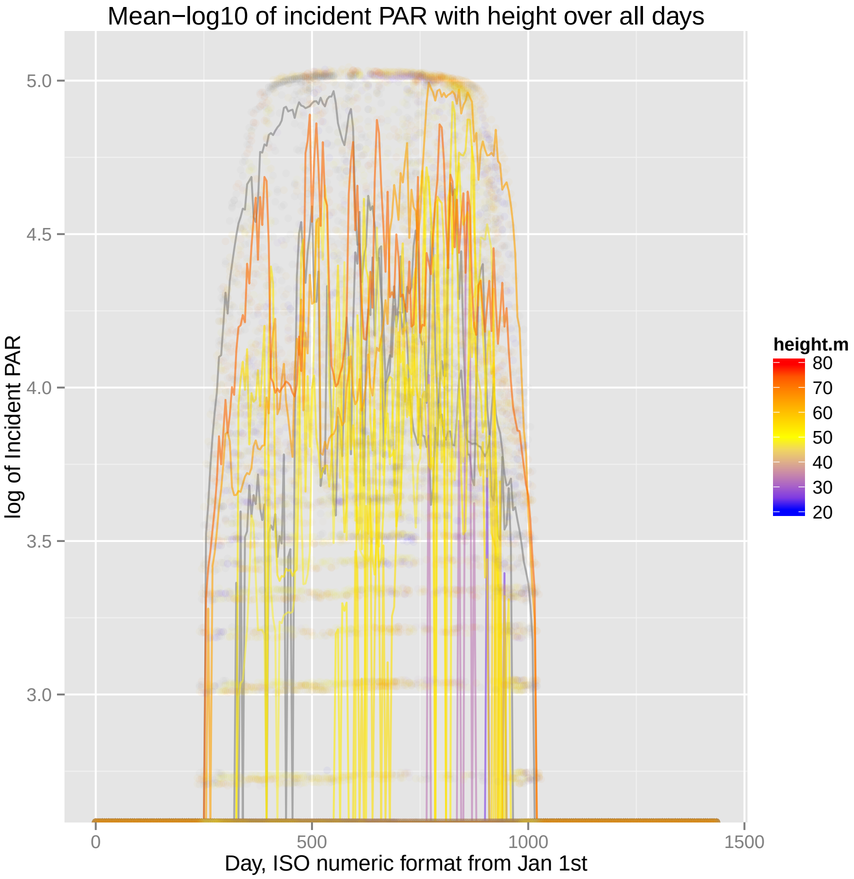

Figure 2: Incident light in time, with heights indicated by color. The higher heights are above the lower heights in many cases, but a simple averaging fails to reveal this trend.

as in plot 3(a), the large numbers of 0's in the PAR measurements make the plots difficult to interpret; in this case, almost every box is populated by only outliers. Furthermore, since we already know the data for incident PAR should be bimodal from 3(a), a box plot is innapropriate to capture trends. A box plot in the logarithm may be an improvement, but fails to capture any bimodality.

**Figure 3(d)** This plot removes the mean of all sensors (heights) from the data at all points in time. These distributions are shown to be quite narrow for nearly all readings, but with extremely long tails. These narrow bodies and long tails make the box plots less revealing than they might be. If the mean is of interest (as is likely the case for PAR), then these plots conceal that information. There is also the problem that our heights are not evenly distributed – thus, taking a direct average of the nodes will weight our mean for each time slice towards the region with a higher point density (the top). These plots show us that the trend from 3(c) in the height vs humidity and temperature are not robust to mean removal. Perhaps these effects are not as straight forward as 3(c) presents it.

**Figure 4** This figure plots the variables against the time-of-day for May 1, 2004 (one line per node), then plots the variable against height for 9:35AM. This demonstrates the consistency of our measurements across nodes for humidity and temperature, but reveals the extremely wide flucutations in PAR across time. Thus, the PAR plots include a spatial averaging. This average is a measure of the total amount of light accessible for that tree. It is not clear why PAR is lower in the morning than the evening, perhaps this comes from foggy mornings. An anlysis across multiple days would be elucidating. It should also be noted that this averaging probably has the problem of being weighted towards more densely populated heights; how to generate a measure of the sunlight accessible by the tree should be considered.

For the height snapshots, they did not include a key describing the pink vs blue points. However, these snapshots are clear otherwise. It is helpful to see raw data for height.

# 4 Findings

## 4.1 First finding

My first finding is that the humidity and the total incident light are correlated. As could be expected, a very foggy day will have both high humidity and low light. What is most surprising about this finding is that the trend does not appear until we nearly saturate humidity. For the interior tree, there is much less of a trend.

It should also be noted that there are few days with mean humidity less than 40%. This coastal region is clearly very humid.

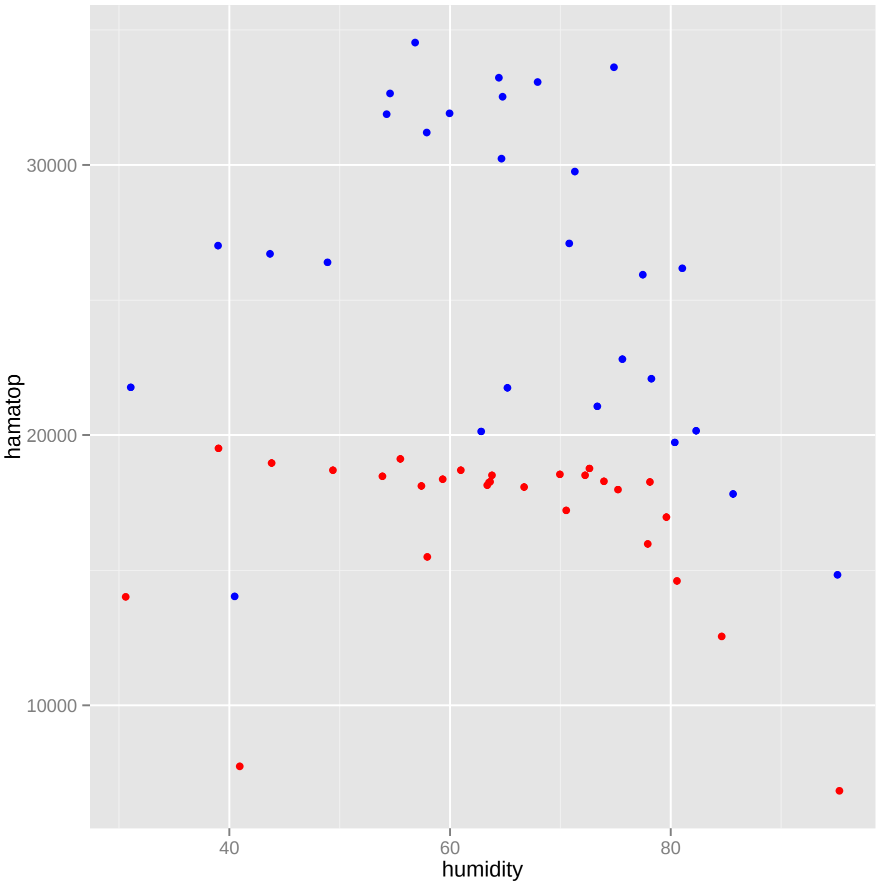

Figure 3: Humidity vs Incident PAR for the edge tree, as averaged between 7AM and 5PM over the full tree. Blue dots label the interior tree and red label the exterior tree. There is a correlation between these variables, stronger in the edge tree than the interior tree.

## 5 Conclusion

I did not learn much about the redwood tree.

## References

[1] Tolle G. et al *A Macroscope in the Woods*. SenSys05, San Diego, California, USA, November 24, 2005