# Project01

*Jonathan Stuart (3032239913) and Nikhil Sakhamuri (3031843166)*

*3/25/2019*

## Data Collection

This was a case study of a wireless sensor network that recorded 44 days in the life of a 70-meter tall redwood tree from Tuesday, April 27th 2004 to Thursday, June 10th 2004. The readings were taken from a redwood tree located in the Grove of Old Trees in Sonoma, California. Sensors were taking readings every five minutes throughout these 44 days, leading to a total of 1.7 million data points over 33 motes. Each mote contained various sensors that measured four different data values throughout their deployment. The four values of interest were temperature, relative humidity, incident photosynthetically active radiation (PAR), and ambient PAR. After the conclusion of the study and analysis of the obtained data, several lessons were learned and several important takeaways were made. The first lesson that was learned was the extreme importance of being extremely precise when setting up delicate sensors. This is because even a slight variation in the position, angle, etc. of extremely sensitive sensors can cause unwanted variation in the data. This is what happened with the PAR sensors in this study; initially it was thought that fluctuations in the PAR readings were caused by differences in received sunlight caused by blocking foliage. However, it was later determined that the fluctuations were similar on similar days, meaning that the fluctuations were actually caused by slightly different orientations for each light sensor. A second lesson that was learned in the end was the importance of having a better apparatus for catching and handling failures. Of the 1.7 million data points that were supposed to be received, only 820,700 were usable for analysis. Loggers with larger memory space are one option to avoid this. In conclusion, the deployed macroscope of wireless sensors captured the complex environmental interactions of the microclimate surrounding a coastal redwood tree. This study affirmed the existence of spatial gradients in the microclimate around a redwood tree and captured enough data to track the changes in these gradients over time. This data can be extremely useful in helping to validate other biological theories. This can in turn lead to a better grasp of large-scale processes of carbon and water exchange within a forest ecosystem.

All the data was recorded during a 44 day period from Tuesday, April 27th 2004 to Thursday, June 10th, 2004. Measurements were taken every 5 minutes during this period. This came out to a hypothetical total of 50,540 real-world data points per mote, or a total of 1.7 million data points over 33 motes. However, only 820,700 points were usable. . The four variables of interest were temperature, relative humidity, incident photosynthetically active radiation (PAR), and ambient PAR. Temperature and relative humidity were measured because they are necessary variables when attempting to analyze transpiration patterns of redwood trees. Incident PAR provides information about the energy available to plants for photosynthesis and gives insight about the drivers for carbon balance in the forest. Additionally, satellite remote sensing measurements of the reflectance of the land surface can be validated by looking at the ratio of reflected to incident PAR. Initially, the scientists considered measuring total solar radiation and barometric pressure as well. However, both these measurements were decided against. Total solar radiation was left out because it required sensor was overly sensitive and PAR readings gave pretty much the same information. Barometric pressure was excluded because it was concluded there were no appreciable differences in barometric pressure across the height of the tree. The data was collected by deploying a macroscope of wireless sensors across a singular redwood tree. Starting at 15 meters from ground level and going up to 70 meters from ground level, nodes of the macroscope were placed on the tree with 2 meters of spacing between nodes. This spacing ensured that the gradient of each variable could be captured with adequate precision. Most of the sensors were placed on the west side of the tree because the west side had a thicker canopy which would help shield the nodes from direct environmental effects. Additionally, each node was placed between 0.1-1m from the trunk in order to ensure that the sensors would record only the microclimatic trends of the tree rather than the wider environment. In order to account for possible failures in the system to properly record and transmit the data readings, a local data logging system was used. The data logger recorded every reading taken by every query before the readings were passed through the larger network. This local data logging system was used later to analyze the performance of the system as a whole. Sonoma-data-log.csv contains the data retrieved from the data logs after deployment while sonoma-data-net.csv contains only the data retrieved over the wireless network. Sonoma-data-net.csv has a much lower yield than Sonoma-data-log.csv; Sonoma-data-net.csv is missing an entire two-weeks of data, whereas that two-week period is present in Sonoma-data-log.csv.

## Data Cleaning

### Finding and Scaling Inconsistent Data

After viewing the histograms below (See Fig. 1), we notice that the voltage data is the only data that appears to be inconsistent between the net and log data sets. In general, the voltage values in the net data set appear to be about two orders of magnitude larger than the voltage values in the log data set. In order to convert the data to the same range, we used a vectorized operation to divide each of the net voltage values by 100, and then reassigned that scaled column to the original voltage column in the net data set.

---

### Comments on Missing Data

In order to count the number of missing values in each of the datasets, we wrote a function that calls `is.na()` in a vectorized fashion over each of the columns in a given data set and sums the number of affirmative responses. Those sums are then stored in a vector and returned as an atomic vector of integer values. That is, we wrote a function to count the number of missing values in each column and store the sum in a new vector. We found that the net data set had the fewest number of missing values with 4263 observations returning a missing value in each of 5 columns. The log data set had roughly double that amount of missing values with 8270 observations returning a missing value in each of 5 columns. We counted missing values for the humidity, temperature, adjusted humidity and top and bottom PAR values (incident and reflected). When each variable returned a missing value, missing values were also returned for the 4 measured variables. All other variables in the data sets yielded no missing values.

With respect to the time period, we found that the log data initially gave no insight into when the missing values were recorded because the Result Time variable was a uniform value throughout the data set. Using Epoch values as a proxy for time values, we produced a visualization that shed some light on the missing values. We found that after the 2500th (approximately 05/08/2004) epoch up until the 8000th epoch approximately 05/25/2004), the missing values followed a relatively cyclic pattern before establishing a rhythm of constant increase. We also found that the number of net-collected missing values was roughly a third of the number of log-collected missing values during this same time period (See Figure 2).

---

### Incorporating Location Data

After incorporating the location data into the net data file, there are now a total of 15 variables in our `new_net_data` dataframe.

---

### Investigating and Removing Outliers

Examining the quantiles and studying the plots in Figures 3, 4 and 5, we came to the conclusion that we should eliminate humidity values over 100%, as was also mentioned in the Tolle, et al. paper. We also found that temperatures below zero and over 100 were likely outliers. We made these decisions with respect to the removal of outliers given out understanding of Sonoma County and the weather it experiences during the months when the experiment was conducted. We also applied the technique discussed within the Tolle, et al. paper whereby records were removed when that record's voltage fell outside of a predetermined range, even though this method is flawed, as discussed in our findings. This decision was made because it is reasonable to assume that sensors with peaking or failing voltage measurements were incapable of recording accurate data measurements, even if only some of the data measurements are compromised, as discussed in our findings.

## Data Exploration

### Pairwise Scatter Plots

Scatter plots of slected variables can be seen in Figures 6, 7and 8. In Figure 6 examining the relationship between temperature and depth of sensor placement, the plot indicates that as the depth of sensor placement increases, temperature drops. Figure 7 compares incident and reflected PAR in the context of depth of sensor placement. Here, we see that the smaller depth values showed greater correlations between reflected and incident PAR, indicating that sensors at a depth of 1 got the most reflected PAR as increases in incident PAR correlated positively with increases in reflected PAR. Senors at higher depths, however, showed less and less increase in relfected PAR per increase in incident PAR. Finally, in Figure 8 which examines the relationship between temperature and humidity, the plot clearly indicates that as humidity rises, temperature drops.

Fot the time period, we chose the time between epochs 3000 and 10000 as our exploratory data visualization indicated that epochs outside of that range showed uncharacteristic data patterns, especially in the range before epoch 3000.

### Incident PAR

Hamabot and Hamatop are associated with relected and indident PAR, respectively. The paper disucsses how motes were able to measure photosynthetically active radiation both toward the top of the tree and toward the bottom of the tree. Motes placed further toward the top of the tree were able to measure incident PAR, and the radiation measured by these sensors came directly from the atmostpher. Sensors toward the bottom of the tree, however, mainly had access to reflected PAR. In this way, Hamatop is associated with Incident PAR and Hamabot is associated with Reflected PAR.

### Temporal Trends

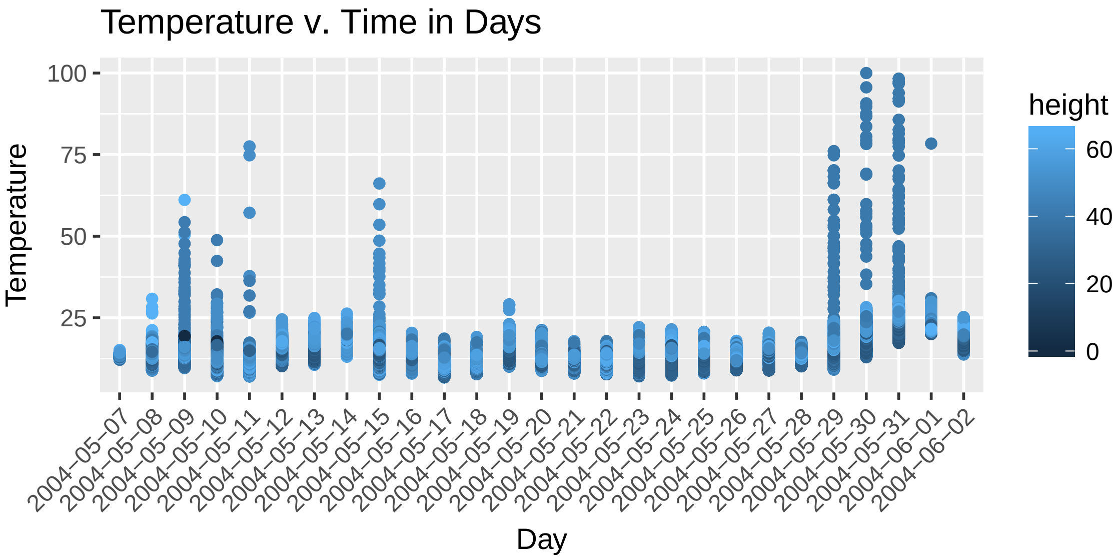

Figure 1: Temporal Trends - Temperature