Lab 1 - Redwood Data, Stat 215A, Fall 2018

Blind

9/13/2018

1 Introduction

In this lab we were requested to analyze data obtained through a network of wireless sensors deployed on redwoods tree within the Sonoma region of California. The experiment was used both for the collection of data regarding the micro-climate surrounding the immediate environment of redwoods tree, and in order to study experimental real-time wireless sensor networks. In this lab, we focus on analyzing the climate data obtained during the experiment. We will hopefully discover new finding about redwoods tree and their environment. As background, the California Coast Redwood (Sequoia sempervirens) can be found all along the Northern California coast region, and it can grow up to a height of 112 meters (National-Park-Service, n.d.). It is able to grow to such heights thanks to the Northern California coast climate and the Pacific Ocean, which provide cool and moist air which keeps the trees continually damp.

2 The Data

2.1 Data Collection

The experiment described in the paper consist of a sensor network of 33 nodes on a single tree. Nevertheless, the data-set consists of data from 72 nodes, and the data in the experiment was in-fact collected from 2 different trees. One tree is is referred to as “edge”, while the other tree is referred to as “interior”. I confirmed this assumption by referencing one of the follow-up publications to the experiment under analysis (Burgess, Pittermann, and Dawson 2006). In the provided data-set, We have no information about the relative locations of these trees in relation to each other. However, the follow up publication provides additional information regarding the locations of the trees. The “Edge” tree was on the extreme western edge of a patch of old growth forest, while the “interior” tree" was within the old-growth forest, but very near the northern edge.

We can use the different trees to identify general climatic behavior of the Sonoma region, but we need to be careful in analyzing data from each tree separately in order to identify correct micro-climate effects. Unfortunately, data from the “edge” tree was successfully obtained only for a small portion of the experiment time: April 27th - May 9th. Hence, I will use only data from the “interior”" tree in order to study micro-climate effect, but I will use data from the “edge” tree in order to confirm macro-environment conditions.

Data was collected using an custom experimental wireless sensor network. Each mote was installed with temperature, relative humidity, and photo-synthetically active radiation (PAR) sensors. PAR is a measurement of wavelengths from 350 to 700 nm. Data from the sensors is logged and stored in a local database on the mote, which includes 512KB of flash memory. Data is also transmitted over a GPRS cellular connection to a network gateway for real-time measurement reporting over the network. Once the local 512 KB flash memory is full, new measurement will not be stored in the local log, but will only be transmitted over the network. In theory, until the local memory has been filled, every measurement that has been transmitted over the network should also be found in a local log. For energy conservation reasons, the network is “awake” for only 4 seconds every 5 minutes. This may affect transmission quality.

Temperature and humidity were measured using the Sensirion SHT11 sensor which provided ± 0.5 °C of temperature measurement error and $\pm 3.5\%$ of relative humidity measurement error. PAR was collected using two Hamamatsu S1087 photo-diodes (on for reflected PAR and one for incident PAR). The sensors were calibrated in two phases: a chamber calibration and a roof calibration. The roof calibration was designed mainly for the PAR sensors, and exposed to PAR sensors to direct sunlight on a roof from 2 days, which

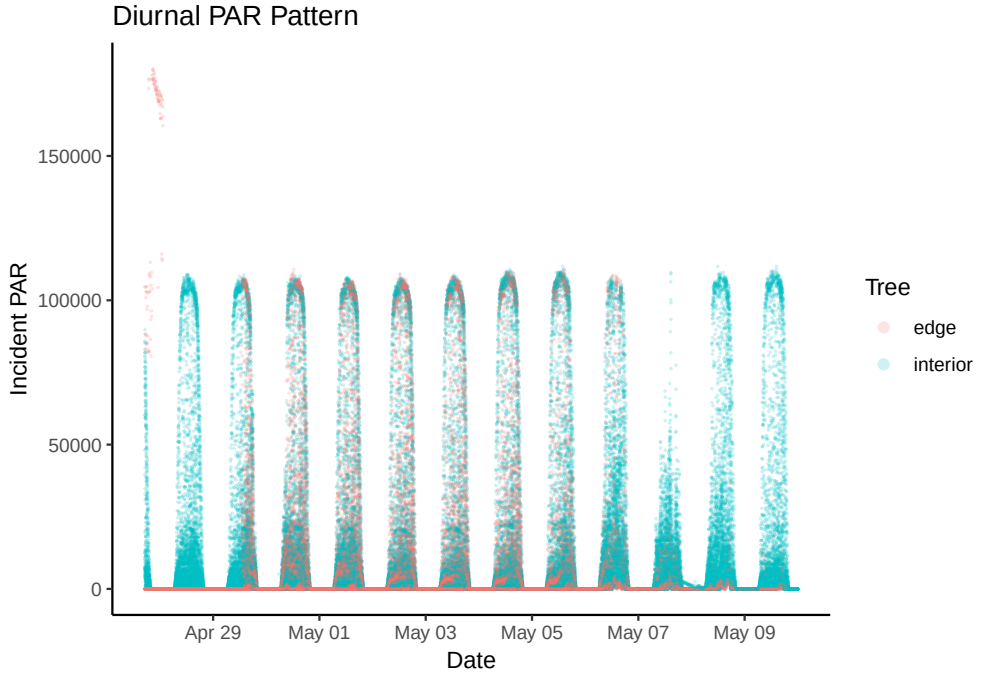
measurements being collected in 30 seconds intervals. These readings were compared the reading from a well-established PAR sensor, and were found to be acceptable. The chamber calibration was designed mainly for the temperature and humidity sensors, and exposed the sensors to a range of temperatures and humidity levels in a weather chamber. The nodes were installed on the deployments site directly from their calibration racks. However, measurement of PAR was still impacted by the deployment location of each sensor node, since each PAR sensor experienced different consistent patterns of light and shade (a result of the surrounding environment) depending on it's installation location and angle.

2.2 Data Cleaning

When examining the data and matching sampling times based on epochs vs. sampling times from the networks time-stamps, I find that there is an offset of 7 hours. This is likely since Pacific Standard Time is GMT+7. Many computerized systems are set to GMT by default, so it is reasonable to assume that the network time-stamps are in GMT, while the epoch time-stamps are in PST. While there is still a mismatch of approximately 2 minutes between the artificial epoch-based time-stamps and the network measured time-stamps, the artificial time-stamps are sufficient given the 5-minute sample resolution of the study.

I also perform basic sanity filtering of humidity under 0%, temperatures below -30°C or above 45°C (which are unlikely in the California coastline). I originally also thought about eliminating humidity values of over 100%, but I found that all the values above 100% were within the humidity sensor's margin of error (3.5%). Furthermore, after further reading I found that humidity over 100% is a possible scenario which is called supersaturation, and it can be measured during fog and precipitation situations, which are possible on the California coastline (especially with the Marine Layer).

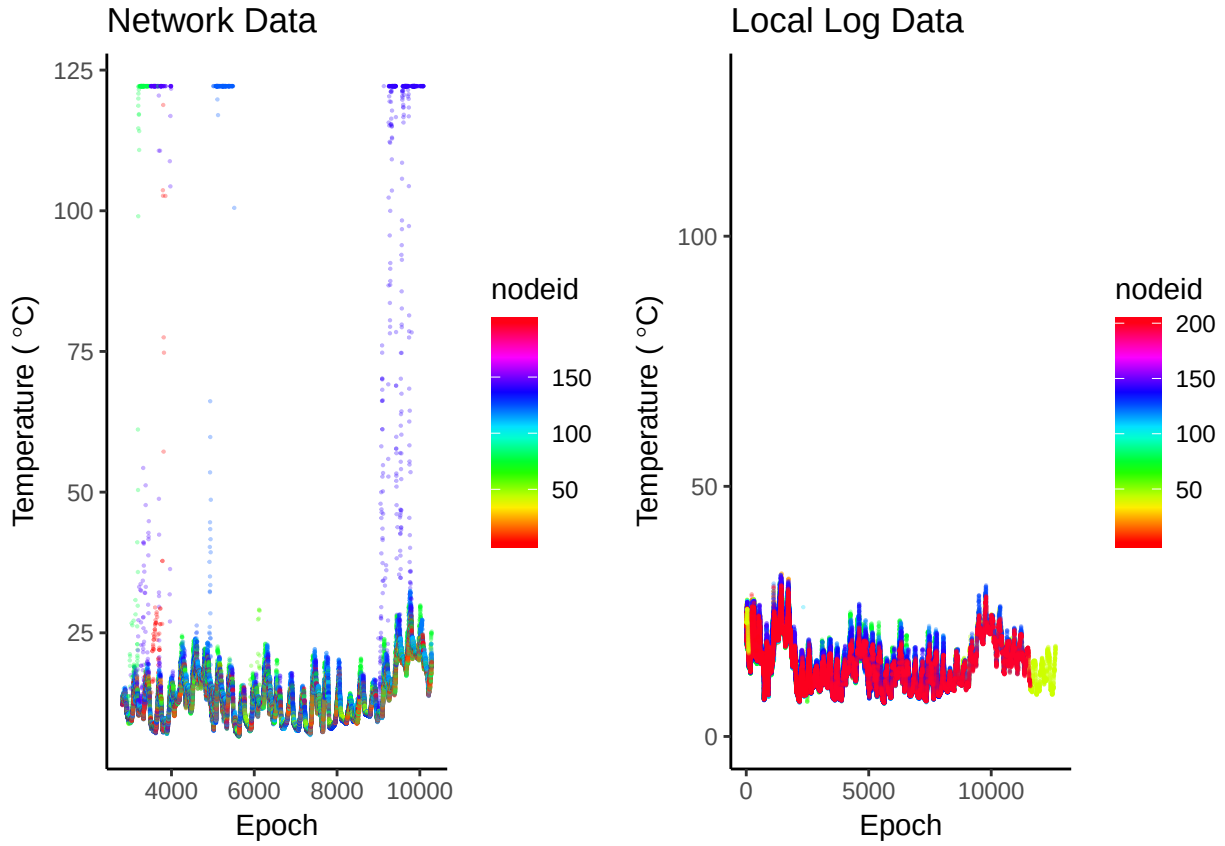
As a final step of sanity-checking measurement values, I wanted to check the PAR values. However, I did not know what are "sane" values for PAR. After plotting direct PAR across several days, the daily day-night cycle pattern became clear, and I could identify several problematic outliers measured during the first day. I found that all the light measurements that came from node 40 were unreliable, and therefore I also eliminated all data from node 40 from the data-set.



I also notice that the measured PAR units are beyond the PAR calibration range indicated in the paper. However, based on the data patterns which match patterns presented in the paper, and based on the diurnal

pattern presented above, I conclude this is likely a scaling factor, and therefore I can continue with my analysis under the assumption of an unidentified scaling factor.

While cleaning the data, I attempted to replicate the paper's findings regarding the association between low battery voltage and temperature anomalies. However, when analyzing the only the local log data, I could not replicate the plot from the paper. When plotting the network data and the combined data, I indeed observe the reported temperature and humidity anomalies and outliers. However, given that the anomalies appear only in the network data (rather than also the local log data), I reach a different conclusion than the original paper authors.



I conclude that while data that was transferred over the network was likely impacted by the battery voltage drop, the data that was measured and stored locally was not impacted. This finding is further strengthened by the fact that many nodes that have low voltage measurements are also associated with no network value at all ("N/A"" measurements). Therefore, for measurements in the "low-voltage"" range (under 2.4 volts), I should reject data that was collected over the network, but keep data that was collected on the local logs. Low battery voltage can result in higher bit-error rate in wireless communication, which can provide a possible explanation of network data anomalies. The outliers in the network-collected data may be a result of data corruption during wireless transmission due to low battery voltage, rather than measurement error in the sensors due to low battery as the original authors assume. Hence, I proceed under the assumption that these outliers are a results of voltage-related transmission data corruption rather than voltage related measurement errors.

I also remove data from sensors with nodeid 100 or 135, since I have no information about these nodes, While they can help characterize general climate characteristic of the macro environment, they will not be able to provide useful information regarding the micro environment. I believe the data-set has sufficient amount of data for the characterization of the macro environment, and therefore I decided to discard the information from these two nodes.

The original paper also mentioned the removal of a node due to poor humidity measurements. I have also identified such a node (nodeid 122). It had significantly different humidity reading compared to sensors at similar heights and directions during the same times. These different reading were exposed during high-humidity situations (especially on April 30th, and May 4th-7th), when the sensor did not reach as high values as it's co-located counterparts.

2.3 Data Exploration

During data exploration, I attempted to explore the general distributions and relationships between different parameters. I explored the variance and dynamic ranges of temperature, humidity and PAR, and the way they are impacted by the spatial and temporal environment. Since the data is highly multi-dimensional, this involved generation of scatter plots across various different dimension combinations while still isolating variables in order to identify correct correlations. The attempts also included fitting curves and focusing on specific date ranges, directions, or heights in an attempt to identify possible spatial or temporal phenomenons. Data exploration was an interative process with data cleaning, since I kept identifying additional suspicious and flawed data-points while exploring the data. These point required further investigation to identify the best method to clean them appropriatly.

3 Graphical Critique

The original paper took the challenge of presenting the multi-dimensional data in various different visual forms. I found the use of different colors to visualize different sensors/heights in figure 4 of the original paper very useful, as it helped emphasize the differences between certain sensors across time. However, a legend would have been useful in order to identify which color belongs to which sensor/node. Furthermore, it difficult to understand the content of the plot without reading the relevant interpretive section in the paper. This is an indication that additional labeling would have been helpful. In addition, given that the sensors were spread across different heights, using a color gradient for different sensors based on their height would have provided even more information in the time-series plots, since it would have been able to convey the relationship between temperature, time and height is a single more intuitive plot. This type of colorized differentiation was indeed used in figure 5 (change in spatial gradient over time), with different assigned dimensions (temperature as color and height on the y-axis). The fitted lines on the right side of figure 4 were helpful to understand the the spatial effects of height on temperature and humidity. The fitted line in Incident PAR spatial plot (third plot on the right side) seemed somewhat in-appropriate due to the large variance and the outlier group of the right side of the plot. The X-axis limits of the Reflected PAR spatial plot (bottom right plot) seemed somewhat inappropriate since there are no plotted values over 50 (??mol/mˆ2/s), which makes it difficult to observe the behavior in the plot

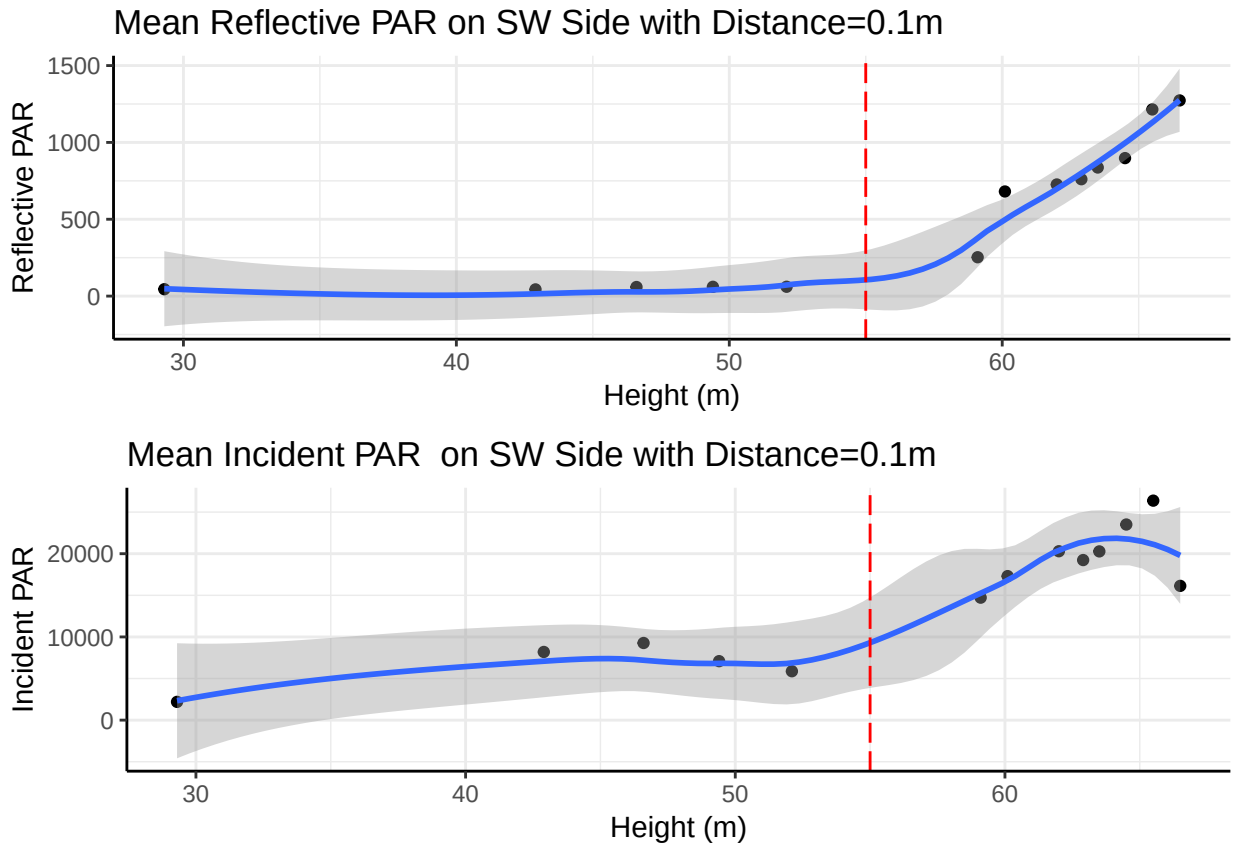
Finally, I found the multidimensional analysis in figure 3 difficult to understand. I contained a large amounts of data, and did not label important trends and outliers. There was no interpretive explanation of the distributions that were presented, and I found part of the information that was plotted in the figure to be irrelevant.

4 Findings

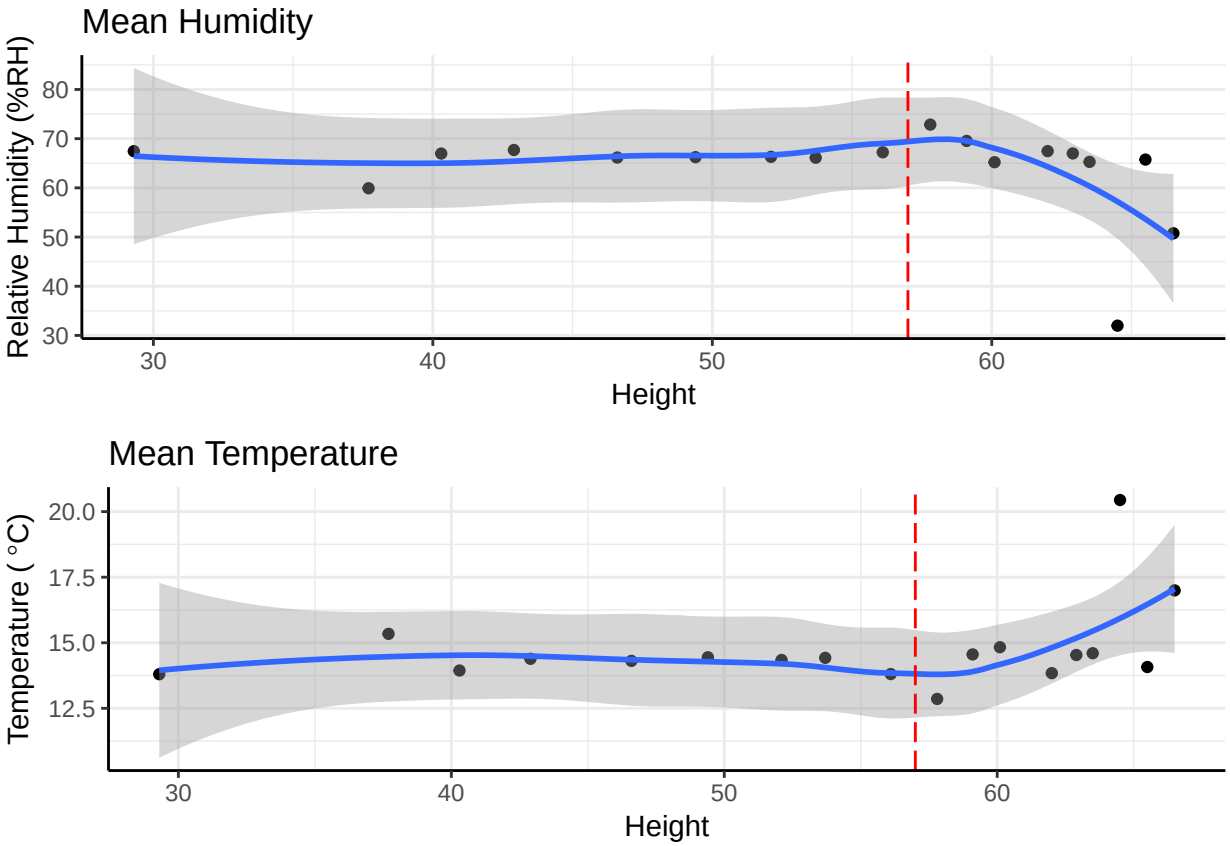
4.1 First Finding: The Crown

The "crown" of the tree, at a height of above 55-57 meters experiences a unique micro-climate as opposed to the rest of the tree. When observing the mean PAR (both Incident and Reflective) across the entire experiment duration, we observe that unsurprisingly, the top of the tree is exposed to more light than the lower parts of the tree. Nevertheless, it is important to remember that all sensors above 55 meters are at a

distance of 0.1 meters from the trunk and installed on the SW direction, while sensors under 55 meters can be at different distances from the trunk and installed in various directions. In order to isolate this effect, we also observe the data only of sensors at different height but at a constant distance of 0.1 meters on the SW side. The trend did not change, and was in fact more clearly observed in the Incident PAR data as well. It is worth noting that the residuals of the fitted curve for this trend are very high when observing incident PAR, as opposed to the low-residual trend of the reflective PAR. We observe that the significant increase in gradient of the reflective PAR occurs at a height of about 55 meters.

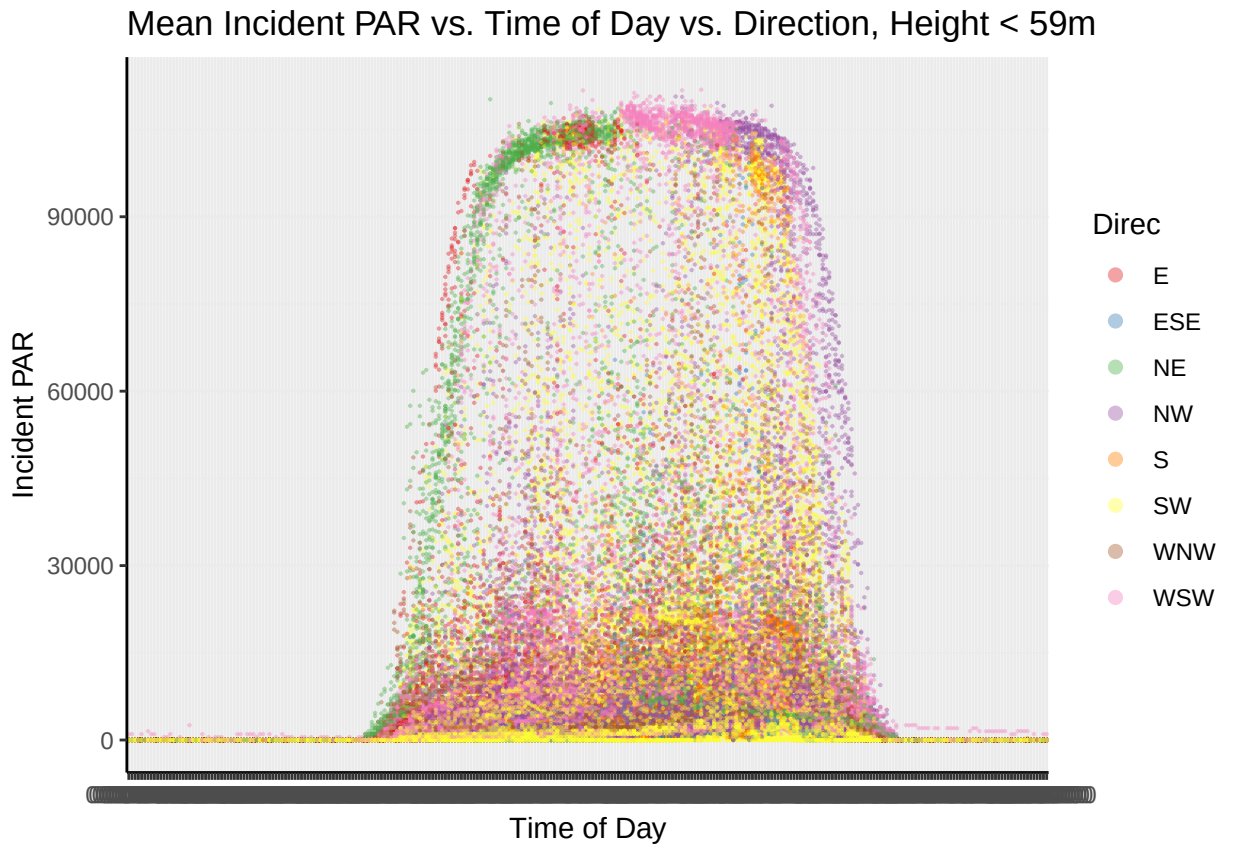


When we take this information under consideration when observing the mean humidity and temperature across the entire experiment range, we can conclude that the top ("crown") region of the tree above is associated with a different climate environment as compared to the lower regions of the tree. While the light exposure data indicates that the crown can be defined at a height of above 55 meters, the temperature and humidity effects are observable only above a height of 57 meters.

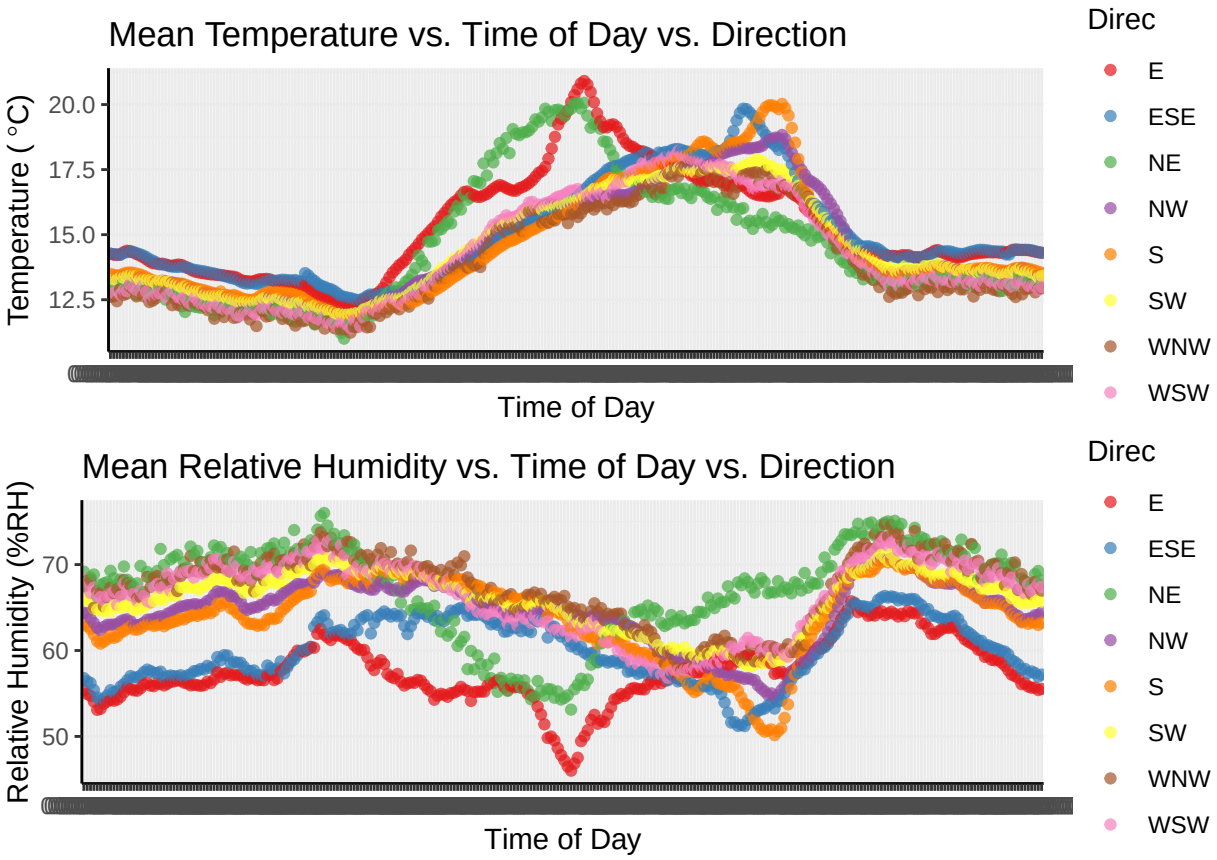


4.2 Second Finding: Sun-Tracking

The movement of the sun during the day hits the tree in different locations, hence creating a gradient of light and climate conditions along different areas of the tree. We can observe that sensors located on the east side of the tree are exposed to more direct light (higher incident PAR) during the morning hours, while sensors located on the south and west sides of the tree are exposed to more light during the afternoon hours. This matches the path of the sun which rises in the east and sets in west, with a slight slant towards the south.

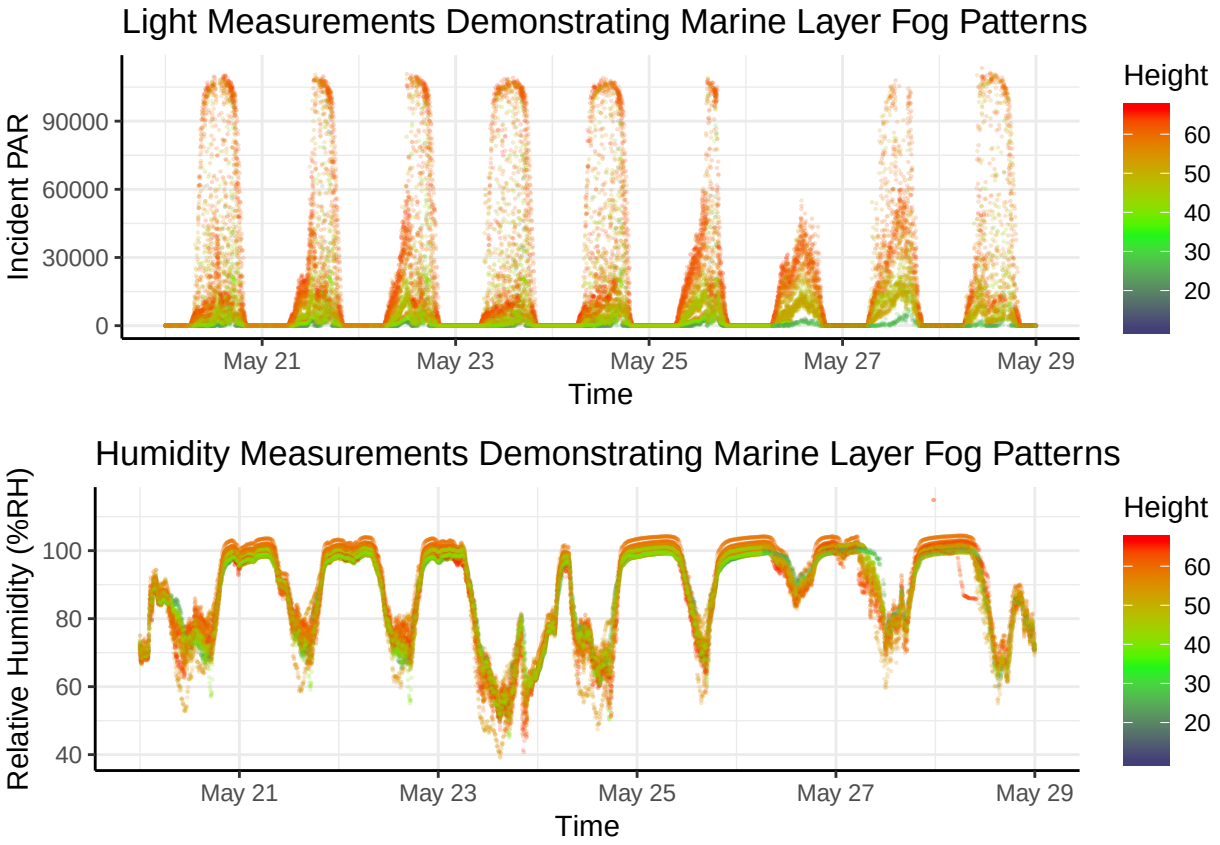


I exclude the "crown" of the tree from the plot, since all sensors at a height of above 59 meters (in the crown) are pointed only in the SW direction. The large amount of light that they receive drowns out the information from all other sensors in different directions. There, observing the data excluding the crown provides a clear image of the resulting light gradients across the different directions in the tree. We can observe a similar relationships in the behavior of the temperature and humidity across different directions: during the morning hours, the west and south sides of the tree experiences higher humidity and lower temperatures, while as time progresses towards the afternoon the east side experiences higher humidity and lower temperatures.



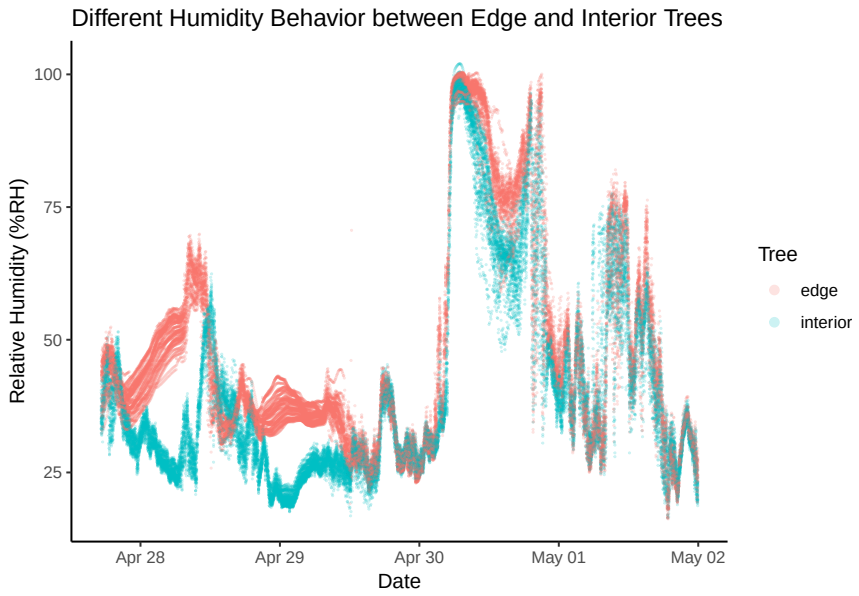
4.3 Third Finding: The Fog

Indicators of the "marine layer" and fog patterns commonly associated with the Northern California coast region during the Spring and Summer months (Wikipedia, n.d.) can be identified using the data. In the marine layer fog pattern, fog and low clouds progress inland from the pacific ocean during the late afternoon, linger overnight until late morning when the ground warms up to evaporate the clouds and clear the marine layer towards the coast. This cycle can repeat for multiple days.This cycle can be identified in the data by observing the high humidity and low incident PAR during the late afternoon and early morning (indicative of fog and low clouds), which give way to lower humidity and more direct light by mid-day. In the figure below, May 21st-22nd and May 25th-May 28th exhibit patterns associated with early summer marine layer fog and cloud. Nevertheless, we can also see that not all days behave the same, and there are also says with plenty of light and lower humidity (such as May 21st, May 23rd-24th and May 29th in the figure below). The high humidity level generated by this fog pattern are part of the climate environment which enabled the redwoods to grow to their unique sizes.



5 Discussion

The little data that we have about the "edge" tree indicated that it has an interestingly different behavior than the "interior" tree. Even though the "edge" tree is supposedly located in a better lit environment, the PAR measurements from it imply worse lighting conditions. It can also be observed in the figure below that while temperature and humidity patterns generally track each other between the "interior" tree and the "edge" tree, there are certain days in which there are significant humidity and temperature differences between the two (April 28th-29th). Additional data from the "edge" tree would have possibly enabled me to further investigate these differences, and the effects of the locations of the tree on their micro-environment measurements.



Also, data about an additional tree would allow for better calibration and outlier analysis of the data, Data from a single tree could be biased for many reasons. The authors of the paper discussed the non-random patterns (which they identified earlier as noise) generated in the PAR sensors as a results of shadows due to sensor locations on the tree. Similar effects could generate similar biased for measurements along the entire tree. Additional measurements from an additional tree would help in eliminating such possible biases. It will be inappropriate to generalize conclusions about redwoods trees resulting from data about only a single tree.

The fact that the majority of sensors were placed on the SW side of the tree made it difficult to analyze the direction dimension of the data, since some directions were represented by only a single sensor, while other directions were represented by multiple sensors across different heights.

Finally, I have learned it is extremely important to have access to the source of the data. Being unable to ask questions regarding measurement values, possible sensor faults, a reasons for measurement error, made cleaning the data-set extremely difficult. Data-set cleaning is difficult enough as it is, so having no domain-expert source to confirm suspicions regarding outliers made it even harder.

6 Conclusion

In this report, I explored the cleaning, exploration and analysis of a raw data-set collected from a wireless network of temperature, humidity and PAR sensors deployed on redwoods trees in Sonoma. I've explored potential reasons for outlier rejection, and I identified 3 findings using the data: The micro-climate at the crown of the tree, the impact of the sun movement during the day on different areas of the tree, and identifying California coastal marine layer fog patterns in the redwood tree environment.

The lack of data-preprocessing an access to the collection source was a major factor in the analysis of this data. Incorrect interpretation of some of my assumptions regarding data filtering could potentially impact the validity of the findings.

Bibliography

Burgess, Stephen SO, Jarmila Pittermann, and Todd E Dawson. 2006. "Hydraulic Efficiency and Safety of Branch Xylem Increases with Height in Sequoia Sempervirens (d. Don) Crowns." *Plant, Cell & Environment*

29 (2). Wiley Online Library: 229–39.

National-Park-Service. n.d. “About the Trees.” https://www.nps.gov/redw/learn/nature/about-the-trees.htm.

Wikipedia. n.d. “San Francisco Fog.” https://en.wikipedia.org/wiki/San_Francisco_fog.