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

9/13/2018

1 Introduction

This report analyzes "A Macroscope in the Redwoods," a research paper from Gilman Tolle et. al. regarding data collected from a single Redwood Tree in California over more than a month of that Redwood's life. In particular, this report looks at the raw data and compares it with the findings presented in the paper to see if they withstand scrutiny. The data presented with the paper was unstructured and messy, and raised many questions about how to remedy problems like inconsistent dates and one variable missing 75% of its values. Once data had been cleaned, the data then presented many questions about the macroscope: If it's real, and if so, how it functions in relation to the wider environment. Ultimately, this reports three biggest takeaways are:

1. Though separate from the larger environment, the microclimate appears to have a relationship with the broader climate's weather patterns given the clusterting and fluctations of relative humidity.
2. At higher humidity, the strong inverse relationship between temperature and humidity holds. This suggests that the demarcation between the macroscope and the general climate at these levels is ambiguous and that the microclimate follows many of same realities as the macroclimate.
3. At lower humidity, an inverse relationship between temperature and humidity holds but it is much weaker. It is unclear if this change can be attributed to the macroscope, but one possible avenue to help determine this would be use edge node sensors throughout the entirety of future studies.

2 The Data

The data analyzed in this report comes from a team of scientists at the University of California, Berkeley, led by Gilman Tolle. They collected the data by installing a wireless sensor network of 72 nodes on a Redwood tree in the Grove of Old Trees in Sonoma, California. They recorded data from this network every five minutes between April 27th and June 10th, 2004 – a total of 44 days. The data was collected by sensor nodes that the team placed every 2 meters between 15m and 70m on the tree. They were placed at both between 0.1m ("interior") and 1.0m ("edge") on the west side of tree. This was done so that the west side's thicker foliage could protect the sensors from broader climatic trends while detecting variation that could exist at different points in a Redwood's specific microclimate.

2.1 Data Collection

In practice, the data collection procedure mentioned above presents a few problems. Firstly, a lot of observations had missing values, though the reasons for these missing values vary. The table below shows the missing values across observations in the raw dataset:

Variable	# of Missing Values
Voltage	301,056
Humidity	12,532
Temperature	12,532
Adj. humidity	12,532
Incident PAR	12,532
Reflected PAR	12,532
Height	6,371

Variable	# of Missing Values
Distance	6,371

The most striking missing values are in the voltage variable, which had 301,056 observations with missing values – roughly 75% of all observations. Upon further investigation, these observations that were recorded at the exact same time – on November 10, 2015 at 2:25pm, many months after the data collection had concluded. It became clear through this process that the scientists added the missing voltage values en masse from flash logs after the fact. This means that, in addition to lacking data on voltage, the result-date variable for all 300,000+ observations was inaccurate. The data would need to be merged with a text file containing all 'epochs', i.e. measurement times, in order for the dates and times to be correct. I describe this processing in the Data Cleaning section.

Next missing values to consider were the five variables (humidity, temperature, Adjusted humidity, Incident PAR, ahd and Reflected PAR) that each had 12,532 missing values. Not suprisingly, these missing values came from the same observations. Zeroing in on these observations suggests that the null values may come from problems with the sensor nodes in certain locations on the Redwood. These observations only come from three nodes. All variables from those observations with non-missing values are within expected ranges. Additionally, every single one of these observations has a distance of 0.1 from the tree. Lastly, all of these observations were on the southwest side of the tree, the heights of their nodes are between 48.4 and 61.1 meters. The missing values here likely have to do with sensor placement. The combination of factors listed above likely made it difficult for some reason for these nodes to pick up data, specifically data collected from the bottom sensing surface of the node.

The last group of missing values to investigate came from Height and Distance variables, which each had 6,371 missing values. These observations also lacked values for direction and location on tree. Like the above values, these ones also come from only three nodes. Within this subgroup, a lot of the readings appear unreliable. For example, the minimum recording for humidity in this subsample was -9,375.37, an obviously impossible humidity measurement that is many magnitudes lower than the minimum sensor value of 16.4. Additionally, the Reflected PAR values seem unlikely to be accurate, with over 75% of them being 0 but the max being 116,000 - well outside the max sensor range of 180. Lastly, the voltage readings are high. While 6,086 of these observations don't have a voltage value, the observations that do all have the exact same high voltage reading of 1,023. While the above nodes' deviations seem to have come from placement on the Redwood, these nodes' clear inaccuracies and large variation for multiple variables suggest sensors that were either faulty or running low on batteries.

2.2 Data Cleaning

2.2.1 Missing Values

Before conducting exploratory analysis, I needed to clean the data so that it would be easily manipulatable. I began by correcting the 301,056 observations that had been labeled with the wrong date and time. This required combining the overall dataset with the sonoma dates text file. This file contained unique ids for each time measurement that corresponded with the epoch variable in the dataset. The sonoma dates file was also, however, highly unstructured. In order to make the data usable, I imported the entire text file as a list of lists. I then used string split to to extract the two lists of interest (Node ID and Date Time.) I then combined the lists into a data frame, coercing the Node ID into an integer variable, and formatted the Date Time via Lubridate so that reflected a more traditional presentation.

After ensuring that the Date Time column had the correct values, I removed many of the missing values mentioned in the "Data Collection" section. I decided this was the best course of action because missing values make it difficult to run data analysis in R. Additionally, the values all came from six total nodes, reflecting less a systemic problem with the data a more a few bad actors creating either no data or, in the case of the nodes that lacked Height and Distance values, noise. Lastly, because the missing values all came

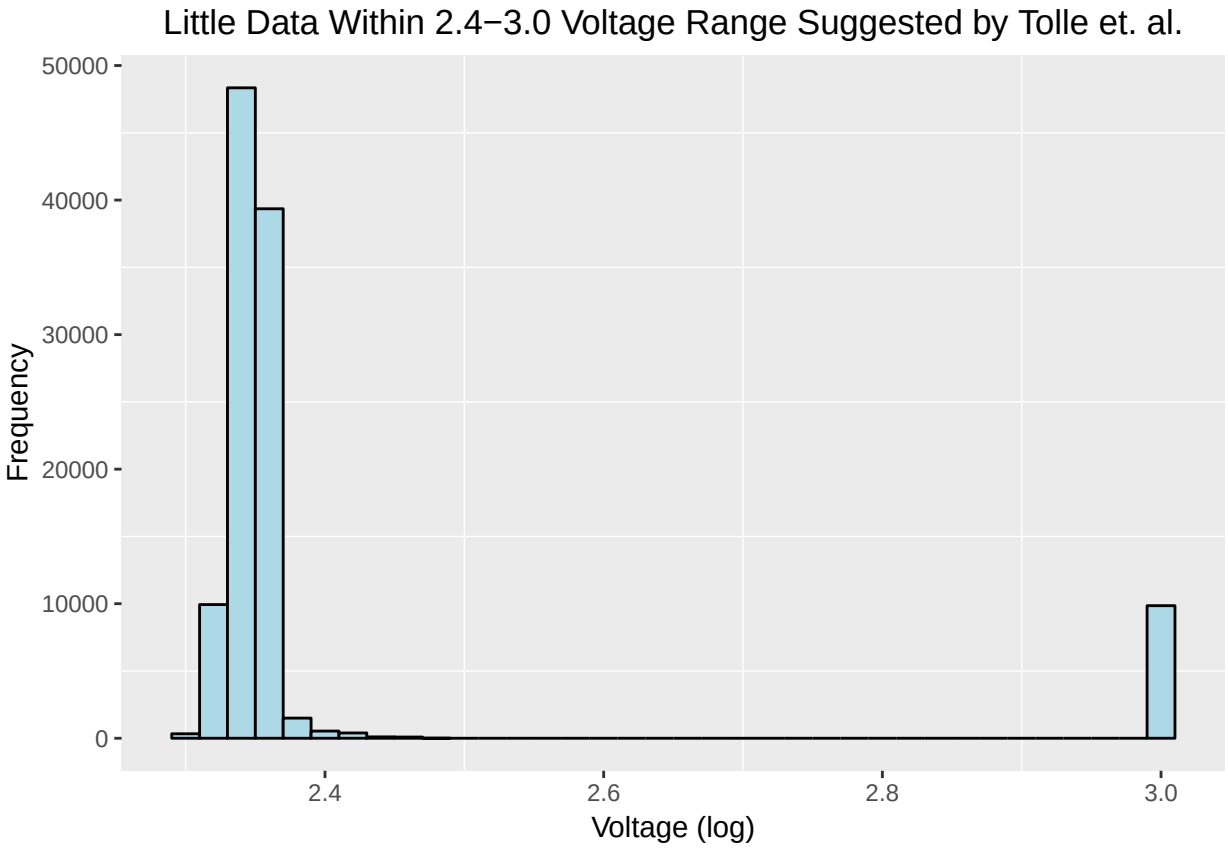
from the same observations, removing them only reduced the data by roughly 18,000 observations, a small fraction of the nearly 400,000 that remained.

One variable whose null values I did not remove was Voltage. This is because over 286,000 observations – nearly 75% of the data – have a missing value in this column. As such, I instead decided to drop the Voltage variable entirely. Removing it now means that none of the observations in the dataset have missing values.

2.2.2 Outliers

After accounting for missing values, I began to identify outliers and eliminate them from the dataset in order to reduce noise and create a more dependable analysis. As the "Outlier Rejection" section of Tolle's paper discusses, the scientists that gathered this data removed tobservations where the voltage fell either above 3 volts or below 2.4 volts due to the correlation between nodes operating outside of this range and various outliers throughout the data Tolle et. al. (2005). One large problem using voltage presents is that nearly three-fourths of the data does not have a voltage measurement, making this a questionable way to remove all outliers in the set.

Additionally, upon reconfiguring the data, it becomes clear that this claim about voltage does not hold up to scrutiny. Though the data as given measures voltage in much larger units, taking the common logarithm of voltage puts it on the correct scale and allows us to judge the scientists's claim. As the histogram below shows, almost all of the data lies outside the bounds defined by the researchers:



Because the vast majority of the data lies below 2.4 voltage on a logarithmic scale, I do not believe that following Tolle et. al.'s automatic outlier rejection range is a good idea. I did, however, decide to heed their advice regarding high end values given the huge gulf between them and the data. This is even clearer on a non-logarithmic scale, where those observations all have a voltage of 1,023, compared to less than 300 for the rest of the observations.

An additional problem with relying solely on automatic outlier rejection is it misses easy opportunities to clean up data that is obviously misreported. For example, there exist observations that lie outside the range of what the sensors could record. The results suggest faulty sensors, meaning the data received from them should not be trusted. The table belows sums up data taken from Tolle et. al.'s paper, as well as my calculations of how many observations lie outside those established ranges.

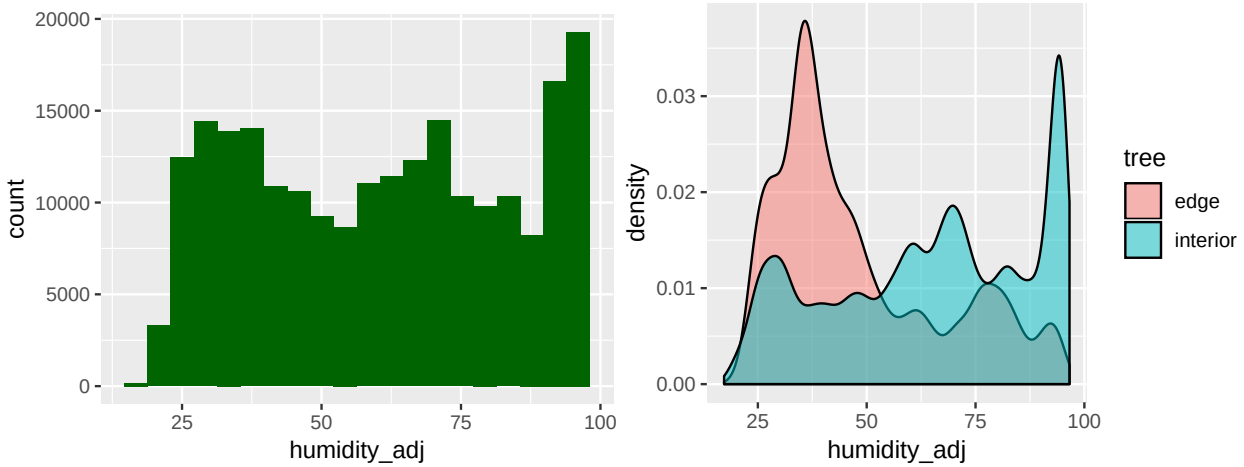
Variable	Minimum	Maximum	# of Observations Outside Range
Temperature	6.6	32.6	1,556
Humidity	16.4	100.2	18,756
Incident PAR	0	2,154	141,997
Reflected PAR	0	180	7,029

Unfortunately, removing the data that appeared outside of sensor ranges means eliminating nearly 170,000 observations, greatly reducing the size of the observations we can analyze. Eliminating a chunk of your data is small price to pay, however, it means the ensuing data can be trustworthy and based on accurate observations.

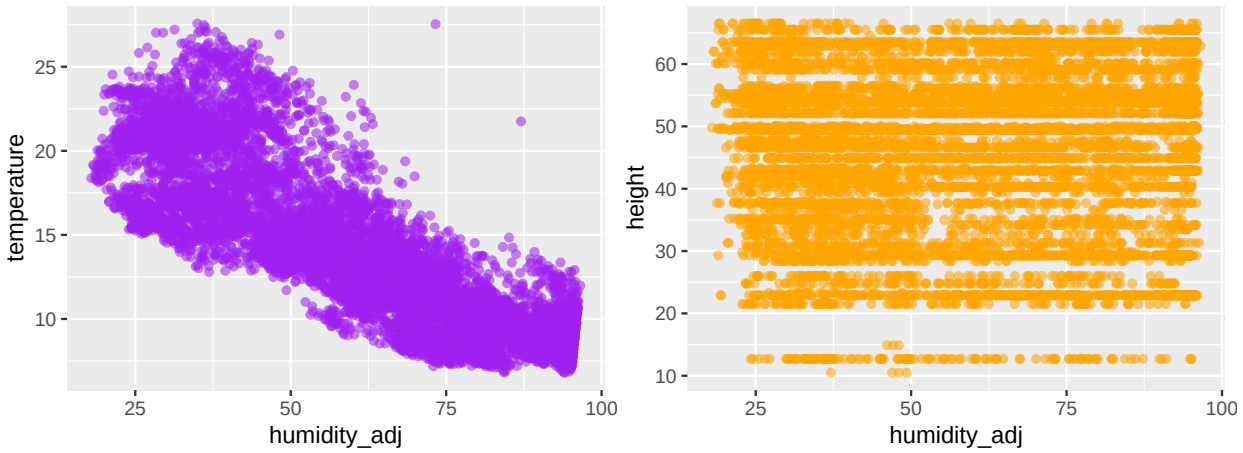
At this point, I have removed all outliers and null values and am ready to begin digging into the data. The last things I did as a part of my data cleaning was remove the result-time variable (given the better Date Time created from combining with the sonoma dates text file), rename variables to more sensible names that better reflected their use, and create a new "Date" and "Time" column to make analysis and plotting easier.

3 Data Exploration

I began by exploring humidity, which I knew would be central to my analysis. Though I was confident that focusing on humidity would produce compelling insights, my interest arose outside of that. I imagined a Redwood Tree in Sonoma County, and one of the first things to come to mind - as with anything in the Bay Area - is fog. I find fog fascinating and the humidity inherent in it the perfect to place to began exploring this data. Thus, I began producing a few quick and dirty graphs to understand humidity. Though my graphs are focused around humidity, they incoporate multiple variables, both here and in my findings below.

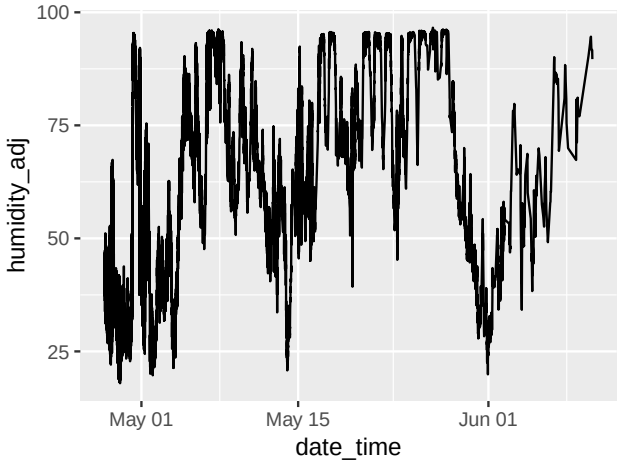


The above histogram and density plot helped me get a better sense of the distribution of humidity. As they show, humidity has a unimodal distribution around the max measurement of 100. This makes a lot of sense given the location of the tree near the coast. Additionally, the density plot highlights the difference between edge and interior nodes, with edges having a much lower relative humidity on average. Though not included in a graph here, a further investgation into edge and interior nodes shows that edge nodes were only recorded at the beginning of the data collection period – a fact I return to in the Findings section.



I furthered explored relationships between humidity and other variables in data via the above graphs. The left graph shows a strong negative correlation between humidity and temperature. This it to be expected: As Tolle states, "One would normally expect relative humidity to move inversely with temperature, because cooler air can hold less water and therefore presents a higher %RH for the same absolute amount of water present in the air." Tolle et. al. (2005). The data backs up this scentific expectation here. The relationship between humidity and temperature is discussed at lengh later in the paper.

To the right of the humidity/temperature graph, we see humidity plotted against the height of tree. There appears to be virtually no relationship between the two. In the interest of time, I did not delve into this relationship much further, but it's good to know that these variables don't appear to have any significant relationship because the focus can turn to other variables.



Lastly, I explored the relationship between humidity and time. As the graph shows, humidity varies significantly over time, going from nearly 100% to below 25%. Interestingly, the humidity isn't immediately oscillating like a seismograph during an eartquake, but instead seems to go through waves where it will be at a higher general humidity for an extended period before dropping to a lower humidity for an extended period. This piqued my interest: Why would humidity stay at a higher or lower general level for certain periods? How long were these periods lasting? I decided to focus my first finding around these oscillations to see if they can reveal anymore about humidity in the macroscope.

4 Graphical Critique

In Figures 3 & 4 of the accompaying paper, Tolle et. al. provide a series of graphs to help the audience understand the data. In Figure 3, they focus on four important variables: temperature, relative humidity,

incident PAR, and reflected PAR. They present each variable in four different ways: histogram, boxplots grouped by day, and two different boxplots based on node height, one of which highlights each box's difference from the mean. These series of graphs want to give an audience a feel for the data. The first two rows help the audience familiarize themselves with the distribution of data generally, while the last two help the audience understand how the data varies at different parts of the tree.

By shrinking these graphs and placing them side-by-side, the scientists raise a lot of questions about the relationships between these four variables. For example, why do temperature and relative humidity have relatively even distributions, while both incident and reflected PAR have a strong positive skew? Why does relative humidity appear to have more variance as time progresses compared to temperature? Why do the distributions for the PAR variables look so different from the distrbutions for variables like temperature and humidity? Why does reflected PAR only have a wealth of observations for high node heights? Questions like these arise for the audience by visually comparing different variables. Some of them are answered visually and through intuition (i.e. reflected PAR has variables at height because it measures reflected light) while some are answered in the paper, but overall this section serves as a "sanity check" - a way for both author and audience to be certain that the data collected generally makes sense and falls within expected ranges. Additionally, this type of thinking about the relation between variables helps the reader digest the "macroscope," and better understand the overarching trends that take place in the life of an organism as complex as a Redwood.

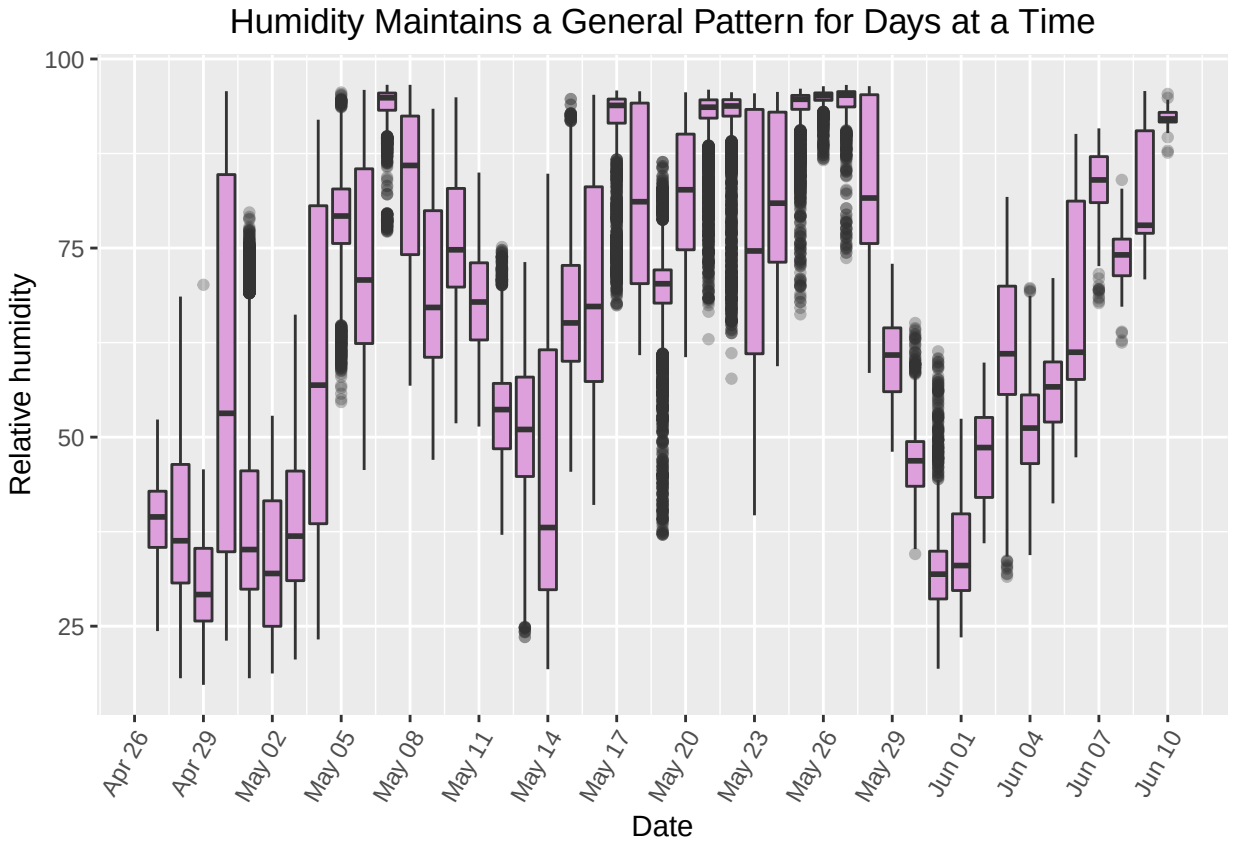
Figure 4 adds to the complexity by focusing on a single day and showing how much variation can happen within 24 hours. As opposed to Figure 3, which served to establish macro trends over the entire timespan, Figure 4 establishes the micro trends that occur every single day. By removing any mention of height on the left side, the graphs can convey the impact the world has on the Redwood. For example, the two PAR time series (and to a lesser extent the temperature time series) clearly highlight the impact of the Sun on the Redwood's microclimate. Moreover, this section shows unexpected results that the authors can assert is caused by the unique of the Redwood's climate. For example, we see that the relationship between humidity and temperature is different than expected. While one normally expects an inverse relationship, on the Redwood we witness major spikes in humidity without any corresponding drop in temperature. This suggests changes in the local climate, further emphasizing some of the main points of the paper.

While I find these series of graphs informative and useful for establishing the author's narrative, I do think the two figures raises contradictions around height. Figure 3 emphasizes the impact of height on readings of variables like temperature, while Figure 4 attempts to downplay that variation and inside emphasize that temperature at all heights moves similarly. Instead of ignoring height on the left side of Figure 4, I would instead have emphasized it. By turning the lines of the time series graphs into colors based on their height, the reader can better differentiate the small but important differences in temperature between the bottom and the top of the tree, while also simultaneously understanding that a rising sun lifts all temperatures.

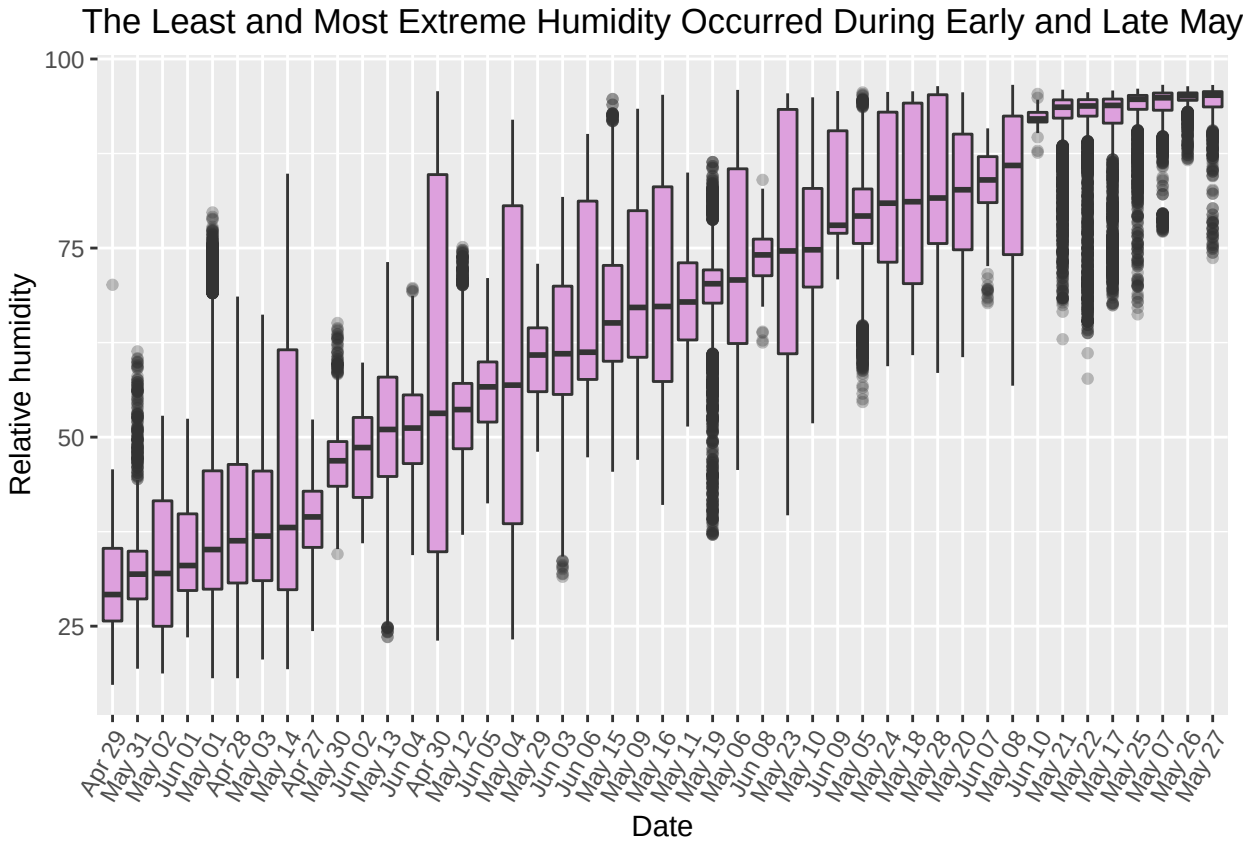
5 Findings

5.1 First finding: Humidity within the microclimate has a relationship with overarching weather patterns

My first finding deals with the variation in humidity across time. To begin this analysis, I created a boxplot of humidity by day, building upon the initial time series analysis I conducted in the Data Exploration section:



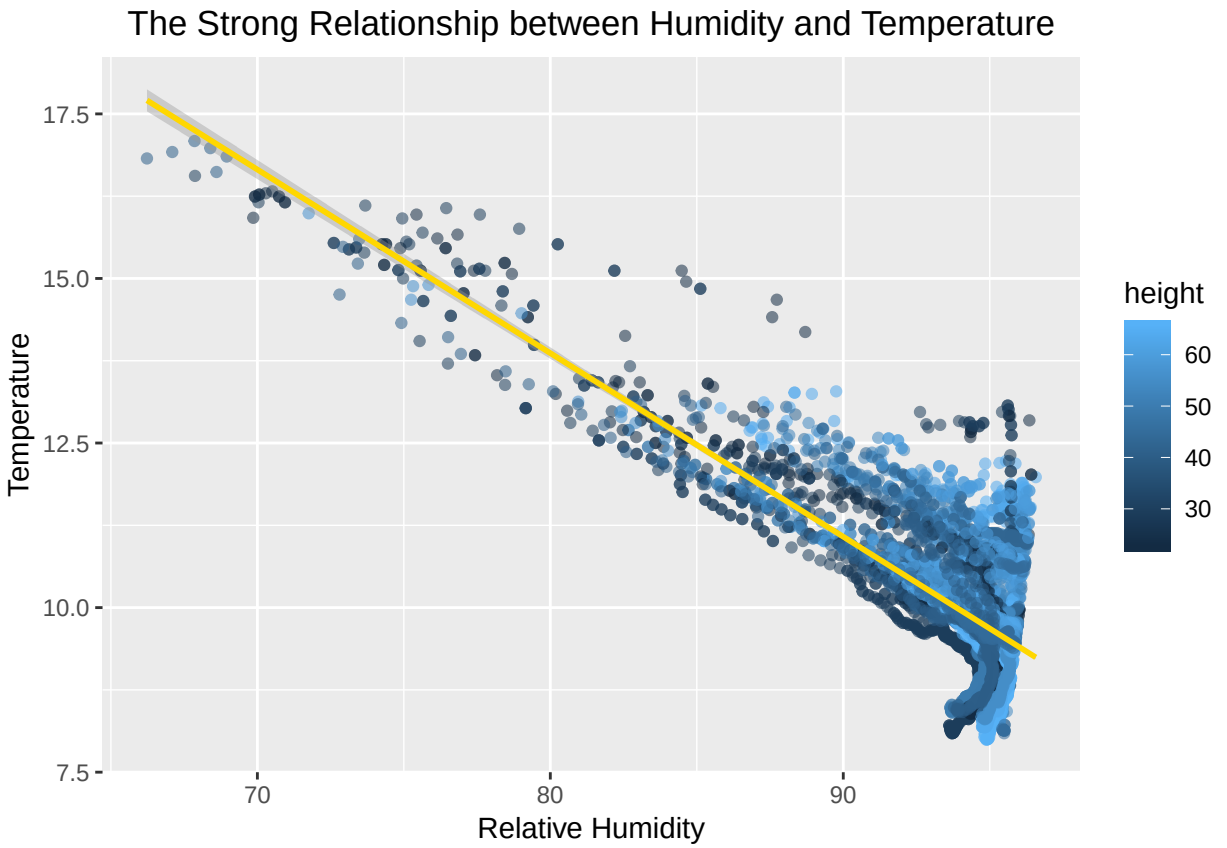
Transforming a simple line graph into a boxplot grouped by day offers a lot of new information for the audience. Now, one can view not only the general pattern the line graph offered, but also see the mean, quartiles, and range for each day. By providing this information, the boxplots help put the initial pattern in better context. As the graph demonstrates, the basic pattern of the same broad humidity for days at a time still appears. At the same time, there are sudden jolts in humidity, such as the jump between May 14th and 15th. The clear change after an extended period of relative stagnation suggests that perhaps weather patterns are changing, thus causing differences in humidity. As currently constructed, howver, the boxplot makes it difficult to further investigate these changes. The switching patterns makes it difficult to read, and it's tough to compare days with similar means that are far apart on the graph. In order to remedy these problems, I reordered the boxplot by median to make it easier and cleaner to view the different levels of humidity across the entire month.



Though the lack of chronological order means a more cluttered x axis, this new box plot is otherwise much easier to comprehend given the natural progression of ascending medians. Once again, the pattern holds: Across the axis, you see groups of dates that are close to each other both temporally and in terms of average humidity. The first group of interest is the on the right, with the highest average humidity. Three of the highest four average humidities occurred within a three-day span (May 15-May 27). While the lower end isn't quite as clustered together, April 27-29 all have daily means within the bottom ten. I decided to isolate these six sets of dates into two different datasets and reinvestigate the relationship I first investigated in Data Exploration in order to see what impact, if any, having particularly high or particularly humidity had on other variables in the data.

5.2 Second finding - The strong relationship between temperature and humidity holds for higher humidity

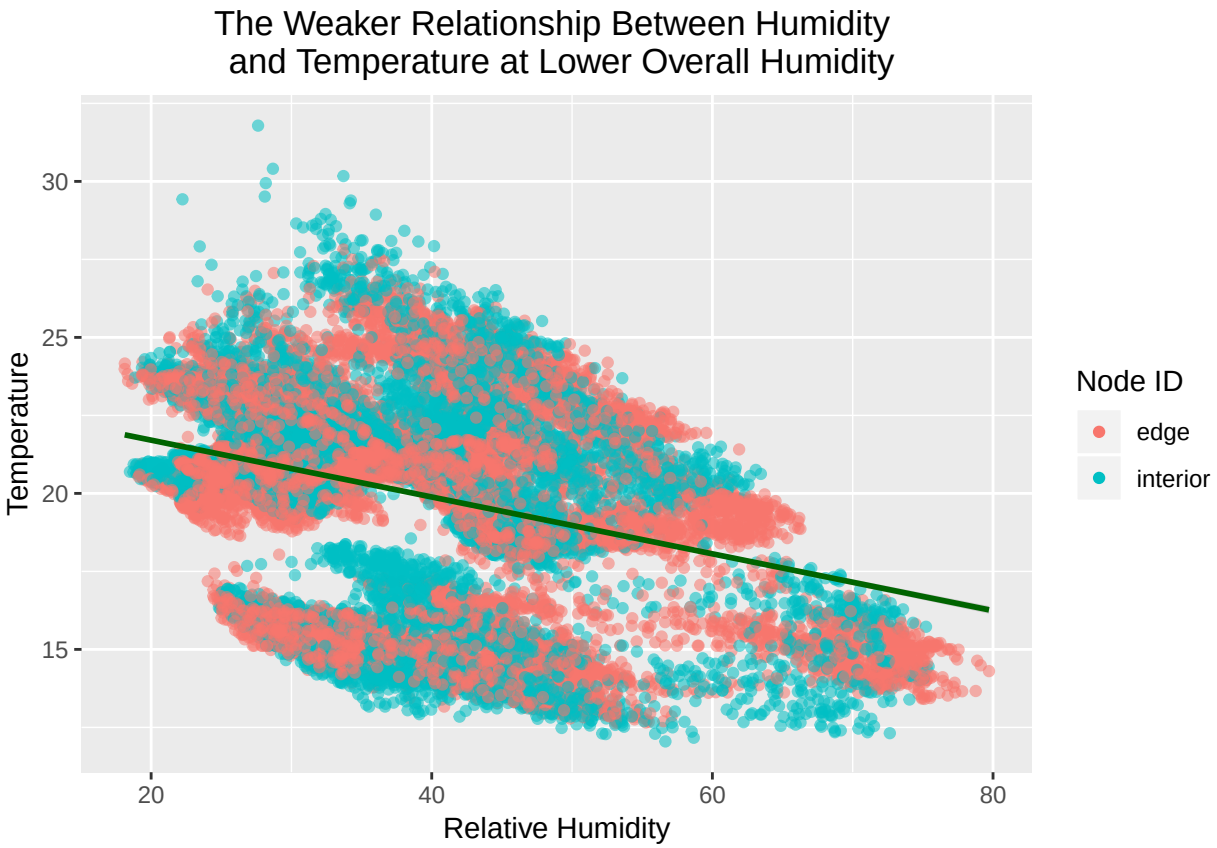
To begin, I plotted temperature against relative humidity for the higher humidity data. If there were no change in the general trends that appeared in overall dataset, then this plot should produce a very strong inverse realtionship, where increasing humidity leads to decreasing temperature and vice versa. While plotting, I provided a least squares line, not to provide regression estimates but rather to provide an easy visual tool for the reader to judge the relationship. Additionally, I include height as a distinguishing variable to see if there were any noticeable variation, but there was not.



Sure enough, the strong negative relationship holds. On the face, this information seems to go against the idea of the Redwood microclimate. If a Redwood has its own, independent climate, why does it have the same basic natural relationships as the environment surrounding it? Tolle offers an explanation, stating that "every microclimatic trend has its own exceptions." Tolle et. al. (2005) Basically, these results are accurate but anomalous compared the overall climate, where one can expect changes in temperature or humidity with concomitant inverse changes for the other. If this claim were true, perhaps we will witness different trends at lower overall humidity.

5.3 Third finding - For data that includes both edge and interior (i.e. low humidity), however, the relationship between humidity and temperature is much less clear

In order to analyze this possibility, I created virtually the exact same plot for the three consecutives dates with lower humidity (4/27-4/29.) One big difference with this plot is the point color: Instead of using height, which offered little information of use for the last graph, I chose instead to use Node ID. Node ID would have also been useless for the last graph because the scientists only used edge node sensors at the beginning of the study (i.e. when humidity was low.)



Though an inverse relationship continues, it is much less pronounced than for the group with higher humidity. Additionally, if interior nodes were removed, the relationship would become slightly more positive. This finding does not dismiss the possibility that perhaps the macroscope does indeed cause a different climate for the tree than the surrounding environment. Moreover, this finding brings into the question the role the edge node sensors play. In the future, I would recommend any study to fully implement and collect data from edge node sensors for the entirety of the study. This will help determine if the placement of the edge node sensors – or some other reason – helped contribute to a less negative realtionship, or if this finding is an anomalous coincidence.

6 Conclusion

Overall, one of the most difficult parts of analyzing the data was its sheer size. It was impossible to do visual exploratory data analysis, as the wealth of points just filled up the entire plot, no matter which plot you choose. My wokraround for this was to take a random sample of 10,000 observations and look at the relationships for that sample. I also did subset the data to specific days or specific nodes for a few EDA graphs that did not make it into the presentation.

In the end, the relationship between the microclimate and the broader climate appears to be very complex and difficult to quantify. While some analyses suggest that the broader climate has a huge impact on the microclimate, other analyses suggest that they are rather independent. This report attempted to help its reader think through these contradictions, and consider ways going forward in which we can conduct a more definitive analysis.

Bibliography

Tolle et. al., Gilman. 2005. “A Macroscope in the Redwoods.” *SenSys ’05*, November. Association for Computing Machinery, 51–63. doi:10.1145/1098918.1098925.