Lab 1 - Redwood Data, Stat 215A, Fall 2017

Hector Roux de Bézieux

September 13, 2017

1 Introduction

The local climate dynamics along the trunk of a redwood tree are complex and hard to measure. Deploying a wireless sensor network of nodes - several captors in one point- along the height of such a tree is a possible way to explore such dynamics.

2 The Data

2.1 Data Collection

The data set is from Tolle et al.[1], and consists of a set of measurements on a redwood tree in Sonoma, California. The data was collected over a period of 7478 days, with measurements every 5 minutes done on 80 sensors more or less uniformly distributed over the height of the tree. 47 nodes are linked to the edge while the rest are labelled interior, a distinction we will come back to.

We started with three data sets:

1. The first one links the position of the node on the tree to its ID number
2. The second one links the date of the measurements to a numerical value, which will be useful to plot.
3. The third is actually consisted of two data sets. Nodes communicate their measurements through a network and also save them in their internal log so we have a net data set and a log data set. Comparing the two values will prove useful for consistency. Nodes ID numbers and date are provided, linking those data sets to the 2 firsts.

All in all, we have access to the following covariates:

- Date and time
- Node ID
- Node parent ID and depth in the network
- Voltage
- Position in height, distance and direction
- Humidity
- Temperature
- Incident Photosynthetically Active solar Radiation (PAR)
- Reflected PAR.

2.2 Data Cleaning

At first, there is 62660 NAs in the data set. **Removing the rows with Nas** removes 3% of all rows. This filter was kept for further analysis.

Then, we can notice that **several data points have values that are impossible from a physical point of view**. Humidity against temperature was first plotted to check whether the physical abnormalities

where correlated, with the appropriate log transformation for humidity ($humidity- > sign(humidity) \times log10(abs(humidity)+1)$. The extreme values are out of physical bounds for humidity and temperature and represent 0.2% of all rows. On a side note, all those values come from the log data set and have voltage values beneath 2.3291, which is considered below functional levels by the authors of the papers. They are therefore filtered out.

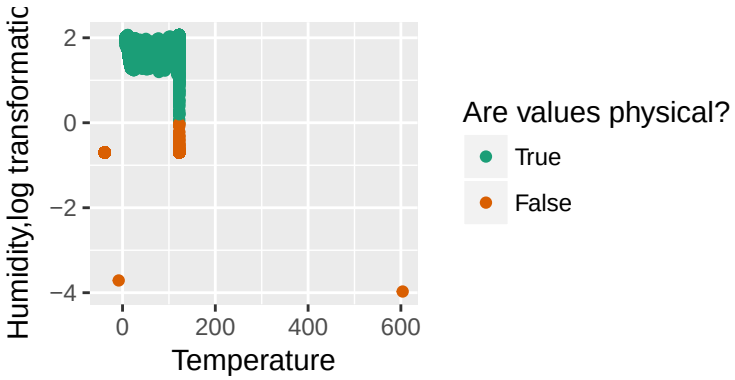

Figure 1: Physical validation of the data

We also looked at Incident PAR versus Reflected PAR.

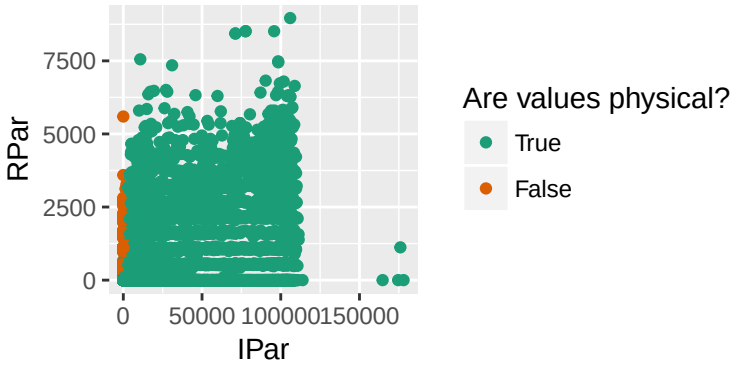

Figure 2: Incident and Reflected PAR

We have a few points with very high IPar. They all come from the log data set and correspond to the same node (number 40). That node is the only one facing south on the edge so we can expect it will receive a lot of light.

Then, 6715 points had a higher reflected Par than Incident Par which isn't possible. For 91.4% of those points, the incident PAR is zero but the reflected PAR isn't. This is probably due to some malfunctioning of the measurement and those rows can be removed.For the rest, it was not possible to find an explanation for the anomalies in measurement but they were still filtered out. As this represent 580 points, it shouldn't have much effect on the analysis.

The remaining data points can be plotted with temperature and humidity:

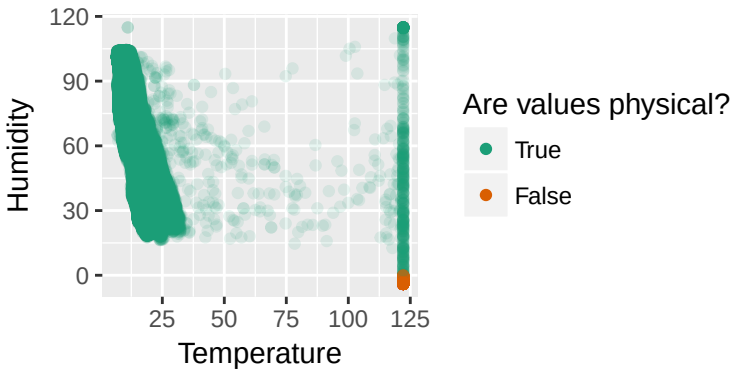

Figure 3: Temperature and humidity after initial cleaning of data

There seems to be a **saturation effect for the temperature captor at** $T \approx 122$. When the temperature captor is saturated, there seems to be measurements errors for the humidity captor as well, and humidity values are all over the place, including some negative values. To stay clear of saturation levels, a threshold of 100 degree was imposed, which filtered an additional 0.2% of the rows.

Then, **the voltage of the nodes** was considered. Here we have very different values for the log data set and the network data set:

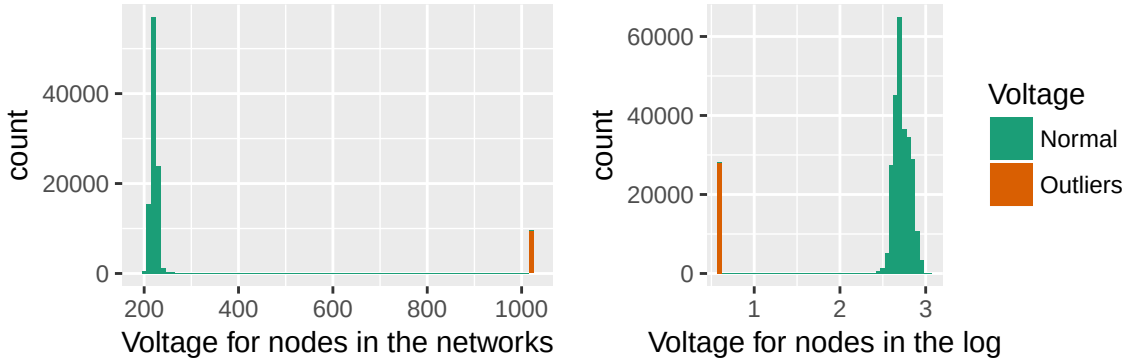

Figure 4: Voltage histograms for the network (left) and the log (right) dataset

All nodes who have a voltage lower than 1 in the log data set have a voltage of more than 1000 in the network data set, and *vis versa*. They probably correspond to nodes with disfunctioning batteries and are therefore filtered. This amounts to filtering 7 nodes. The fact that voltage levels have a two order of magnitude difference between both data sets will be handled in the data exploration section.

2.3 Data Exploration

The data set provided also as a measurement called $humid_{adj}$**.** If the two densities are plotted, it seems like the $humid_{adj}$ values are just a re-scaling of the humidity measurement, probably because humidity is actually relative humidity and as such cannot be over 100%. Therefore, we will now only use the adjusted humidity and refer to it as simply humidity after that point.

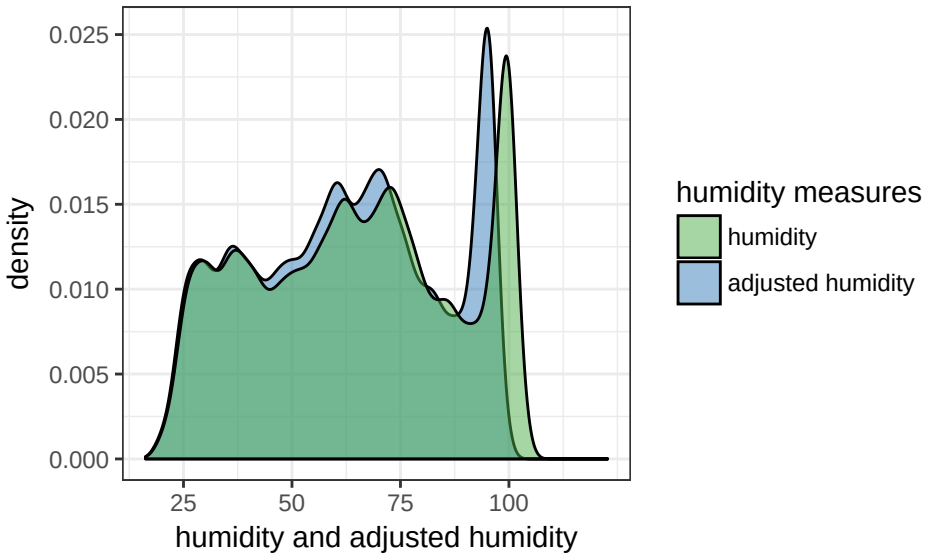

Figure 5: Density plots for humidity and adjusted humidity

To ensure consistency between the two data sets and ensure we had cleaned the data from most abnormal points, we selected the points that were matched in both data sets.This yielded 77965 data points. Throughout all those points, temperatures, humidity levels and Reflected PAR matched perfectly or with a .3% difference at most that we decided to ignore. For the Incident PAR, all matched perfectly except one where the IPar had a three-fold difference between the log and network values. This point was in fact a duplicate in the log data set and appears twice, once with the right IPar value and once with the inflated one. We therefore deleted that point.

As mentioned above, the voltage levels for the log data set range from 2.28431 to 3.0302 while those in the network data set range from 198 to 287. By selecting the common points of both data sets, we can see a clear relationship between the two voltage levels. A linear fit (in red) is not sufficient to fit the data. A better fit is a lin-log model (in blue).

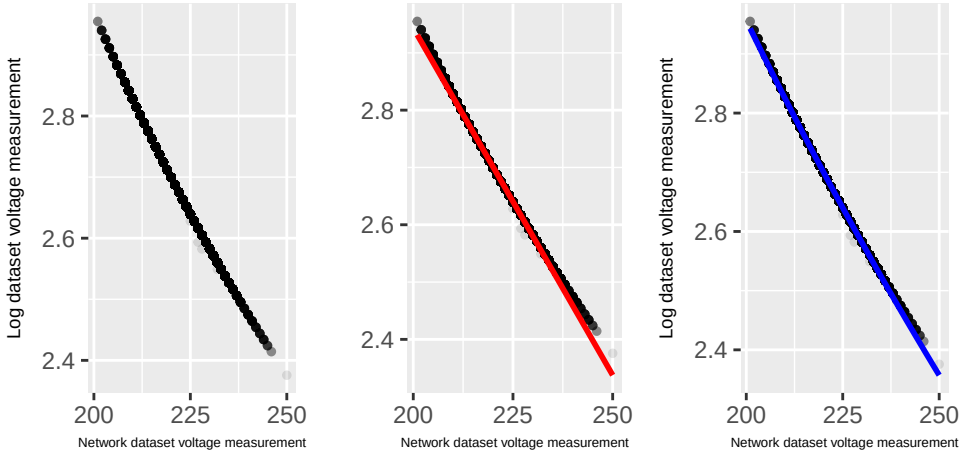

Figure 6: Density plots for humidity and adjusted humidity

We will therefore keep that option and we use that linear model to re-scale all the voltage values to be similar to the log data set.

The next step is then to **check for consistency between the data sets derived from the node measurements and the others**, namely the date and position data set. We use the position data set as a reference for Node ID and filter out any node whose ID doesn't match. There is one node in the redwood that doesn't match any ID in the position data set and 11 from the position data set with no measurements (including some that have already been filtered out).

If we look at the dates data set, we can see it spans 46 days while the network data set only spans 27 days and all the points from the log data set are registered on the same day. Other than that, the date data set did not yield much insights into the data. We can notice that we have a perfect match between the time points from the network data set and the epoch. Therefore, we will ignore the date data set and use epoch as a proxy for the time afterwards.

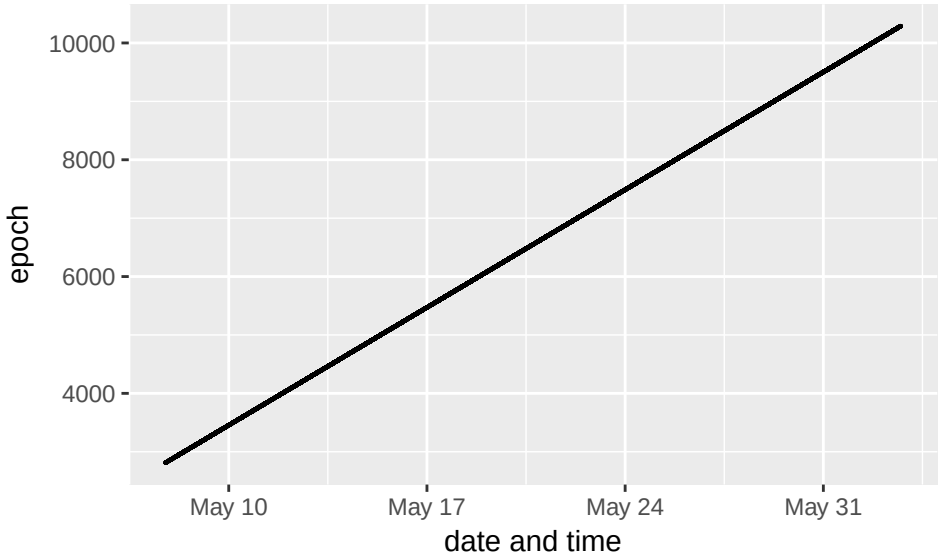

Figure 7: Relation between the date-time and the epoch variable

Finally, the position data set distinguish between two types of trees: interior and edge. The edge nodes are only present in the log data set and only have points for the first 12 days. We will discuss the differences between each tree in deeper details in the *Findings* section.

3 Graphical Critique

3.1 Figure 3

The aim of Figure 3 is to help the reader get a clearer sense of the data. As the authors point out, the challenge is to represent a 3-dimensional data point (time, height and measurement) on a 2-dimensional sheet of paper. To do so, they proceed in three steps:

1. Step 1 is a 1D visualization of each measurement. Those graphs serve their function rather well, even if the number of bins is a bit small and enhance artificially some aspects of the data. In particular, the first mode of the humidity measurement is quite smaller with a smaller bin. However, all in all, Figure 3)a) of the paper fulfill its role of giving a first approach to the data.

2. Step 2 corresponds to Figures 3)b) and 3)c). Those graphs aim to project the data on a 2D plan by squeezing the 3^{rd} dimension in the form of a boxplot. The general idea is pretty convincing and is

appropriate to the data. However, those graphs are pretty messy. The color is very light and the eye is drawn to the outliers who actually represent many more non-white pixels in each plot, especially in 3)b) where we can't even see the boxes for the RPar measurements. Plotting the boxplots with or without the outliers could be one way to solve this issue. Adding a trendline which follow the temporal or spatial average could also be helpful if the authors wanted to limit the number of plots. On a smaller levels, the axis could be simplified by adding less graduations. In 3)c) in particular, a common y-axis could have worked for all 4 plots.

3. Lastly, Fig 3)d) aims to expose constant trends along the height if the tree by normalizing by the temporal mean. As this graph tries to exhibit global trends along the height, adding outliers only serves to confuse the reader. Adding a full line at x=0 would also help to better perceive the repartition above and below the means.

3.2 Figure 4

Figure 4 focuses on a single day and tries to exhibit both common responses and difference as a function of height. The graphs are divided in 2 columns.

(a) Column 1 display all the points for all four measurements. Colors help to identify each node for temperature and humidity. Those graphs are really nice and easy to grasp. The first two aim to demonstrate the common response of all nodes, and show how they move together in temperature and humidity along the day. The 3^{rd} and 4^{th} graphs aim to show how PAR measures are linked to the sum rising and setting. Not displaying individual lines is a appropriate choice. The choice of green to represent light might be a little awkward. Yellow is hard to see on paper but a deep orange could have been more relevant.

(b) The second column is much harder to understand. For a specific time point in the day, the authors were trying to display the variations in each measurement along the height of the tree. However, the color and icon choices make reading the graphs difficult. First, the main feature of the graph is the trend line, as it is what the authors spend the longest commenting. Therefore, it can be surprising to draw that line so thin and in a fainter color that individual points. Moreover, the choice of cones to symbolize gradients is unfortunate. Regular arrows are more intuitive for gradients. Using different colors to pick out points that go against the global trend is nonetheless a good idea and helps to quickly see outliers, the second important feature of the graph.

4 Findings

4.1 Node working time

We first wanted to look at why nodes failed across time. Different nodes had very different duration time. Some failed quite early and some lasted for the whole observation time. The authors of the paper argue that it is because of insufficient initial battery levels but some other factors could be at play. For example, captors at specific height could be less protected from rain and therefore drain their battery faster. Distance from the center of tree might mean more encounters with wildlife that may damage the nodes. However, as can be seen on Figure 8, there is no relationship between height and battery life, or between distance and battery life. In similar fashion (not shown), the lifespan of each node seemed uncorrelated with any other metric. The initial battery level is therefore probably a good hypothesis.

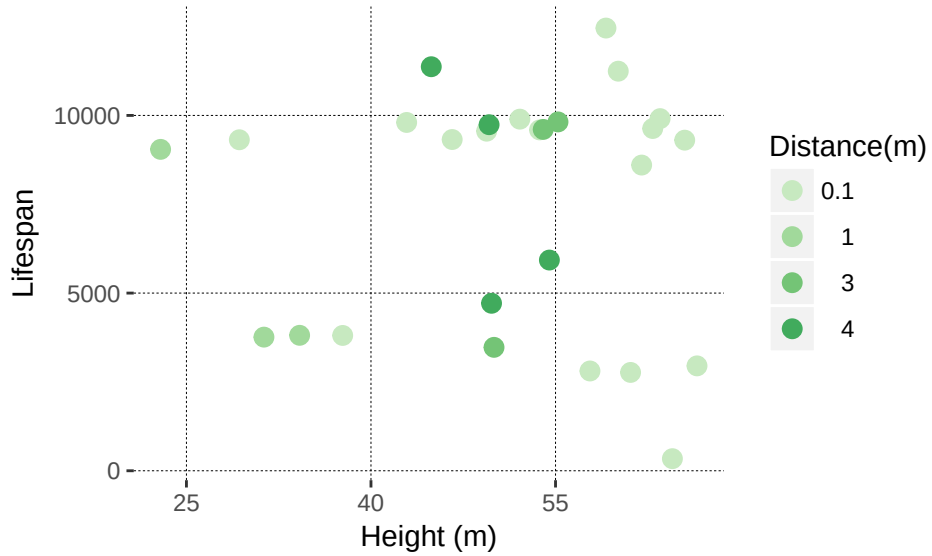

Figure 8: Nodes lifespan according to its position on the tree

4.2 Temperature variations

We then wanted to explore the drivers of temperature variation and more precisely the characteristic distance over which those variations took place. On Figure 9, we can see the variation in temperature for the interior tree over the full observation span. We only kept the measurement that matched between the log and the network data set to avoid any outliers we may have missed in the cleaning process and plotted the temperature of every single node along the tree, colored by the height of the node. It is nearly impossible de distinguish between individual nodes on the plot because the main driver of temperature is the global climate environment of the tree. The main variations in spatial temperature appear at a larger scale than that of a single tree.

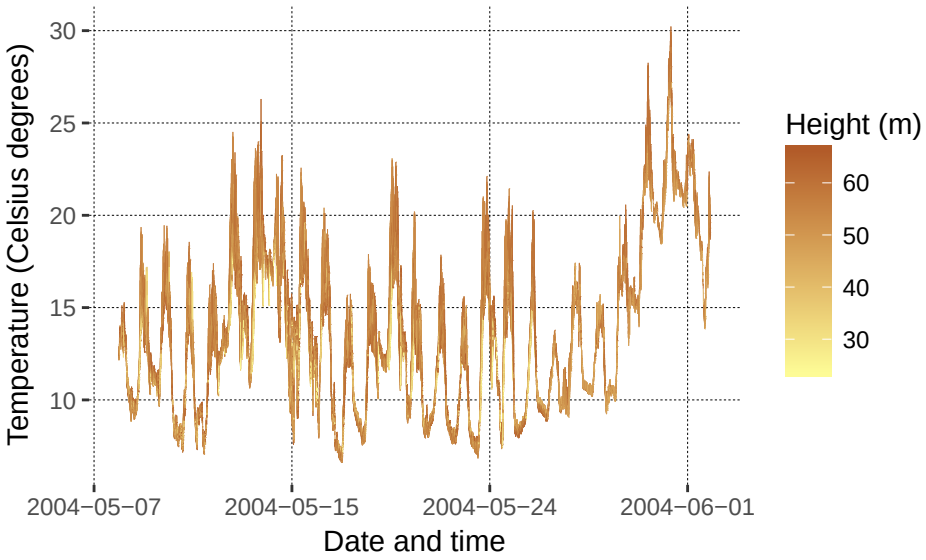

Figure 9: Variations in temperature for nodes of different height in the same tree

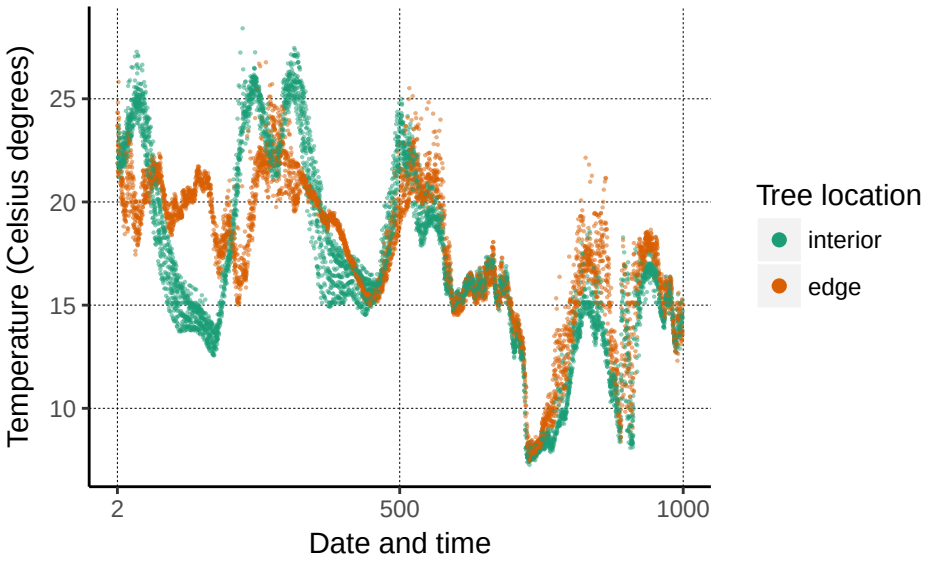

Figure 10: Variation in temperatures for all nodes in two distinct trees

Therefore, we plotted temperature variations for the two trees on a representative time period. As we can see on Figure 10, the temperature variations between the two trees are sometime correlated and sometimes not. This mean that the distance between the two trees is the right scale to study spatial temperature variations. Across the physical span of a forest, temperatures may vary greatly, according to those preliminary results. It also means that a few nodes per tree would suffice to map the forest, as all nodes from the same tree give similar temperatures. Deploying the network in such a way could yield very interesting insights into local climate dynamics.

4.3 Negative correlation of temperature and humidity

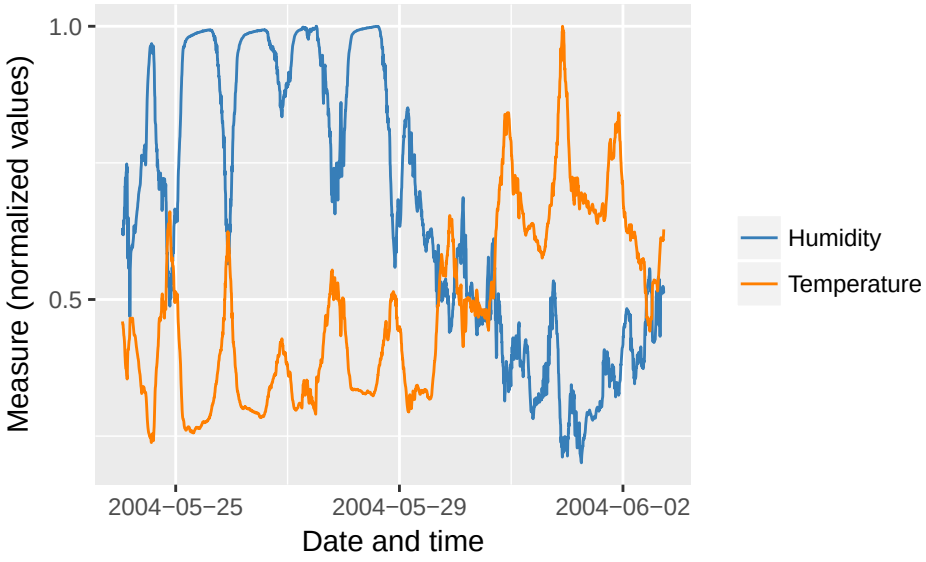

Figure 11: Correlation between humidity and temperature for one node

Finally, we wanted to look at a specific node. We chose node 119 because it had of of the longest lifespans and none if its measurements had been previously filtered. Here, we can look at a representative portion of time (plotting the whole experience time-span lead to unreadable graphs). To get interpretable values, we scaled both humidity and temperature by their highest values to bring everything back to a [0,1] scale.

We can clearly see that humidity and temperature are strongly negatively correlated on the graph. Indeed, on the whole data set, the correlation coefficient is -0.8851974. This could be expected given the shape of the points in Figure 3.

We can also see that the time plotted spans 10.5 days: every day is marked by a local spike in temperature, which confirm that temperatures are more impacted by the passing of time during the day than any spatial metric, for a given tree.

5 Discussion

All the data cleaning performed in this analysis ends up clearing 34.9% of the data set when removing duplicate and 51.3% when only selecting the interior tree, which is roughly what the authors of the paper have, since they only mention data points from one tree. So we can imagine our cleaning was roughly equivalent to theirs. Cleaning impossible physical values and obvious outliers seemed anyway a reasonable thing to do considering the data set and the nature of the outliers.

Our assumption that the variable $humidity_{adj}$ is actually the adjusted humidity may be false and further exploration to re-scale the humidity variable may be necessary in that case. Replacing $humidity_adj$ with humidity did not however alter much our results so this should not be too much of an issue.

The issue of the voltage should be further pursued to insure our analysis is correct. A link with battery lifespan would prove interesting in particular. On that subject, deploying only nodes with full battery, as suggested by the authors, would help develop a better time-series analysis across all nodes.

The question of spatial variation of temperature across the forest could be off interest to climatologists. Using the node technology could allow to precisely map the forest in an efficient way and, paired with terrain information, it could help better understand local temperature dynamics. On the scale of a single tree, the impact of the time of the day is very important. Collecting data points more regularly (every mn for example) on a shorter span of days could help better map the local dynamics along the height of the tree.

Finally, the connected dynamics of each measurement should be studied and, with subject knowledge, a proper model could be developed to explain the relationship between them - and the climate dynamic along the trunk of redwood tree.

6 Conclusion

After a necessary cleaning step that removed around a 3^{rd} of the data points, analyzing the measures yield several interesting results about the heterogeneity of battery lifespan, the spatial and temporal scale of temperature variation and the relationships between metrics. This first analysis gives several clues towards what questions to ask an expert of the field, as well as what a next experiment could look like and what insight might be learned from it.

References

[1] A macroscope in the redwoods, *Tolle, Gilman and Polastre, Joseph and Szewczyk, Robert and Culler, David and Turner, Neil and Tu, Kevin and Burgess, Stephen and Dawson, Todd and Buonadonna, Phil and Gay, David and others*, Proceedings of the 3rd international conference on Embedded networked sensor systems.