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

Kevin Benac

September 14, 2018

## 1 Introduction

The development of wireless sensor networks has enabled scientists to gather a huge amount of data both locally and wirelessly simply by putting sensors on trees. This then allows statisticians to perform various analyses. In this report, we are interested in data on microclimate changes over time of coastal redwood canopy, which we aim to analyze and hence help us better observe environmental characteristics that used to be impossible to measure.

## 2 The Data

We investigate the data collected by Tolle et al. (2005) on redwood trees. According to the original paper, the data was collected from a 70m tall redwood tree in Sonoma (CA), between late April and early June over a period of 44 days, with the help of 33 wireless motes.

### 2.1 Data Collection

A lot of information about the microclimate of the tree has been gathered such as temperature, humidity, incident and reflected photosynthetically active radiation (PAR) data. Each of them has been recorded every 5 minutes, which enables us to analyze the microclimate variations for redwood trees at the end of the spring. Every 5 minutes, the network stays awake for 4 seconds in order to transfer the data and then goes back to sleep in order to save the battery. The motes were put at different places on the tree at heights varying between 10.5m and 66.5m and at different distances from the trunk: between 0.1m and 5m. The data was also recorded locally on the flash chip in case of failure of the network. We therefore have two datasets, a network one and a flash log one.

### 2.2 Data Cleaning

Since we have two datasets we are going to clean each of them separately and then check how we can combine them to avoid duplicating the information. We are not using the `redwood.all` dataset since it is just a stacking of the other two datasets and we believe it is better to clean each of them separately before merging them. We start with the flash log data. This dataset is supposed to be the most accurate one since the network dataset is much noisier (see Figure 7 from Tolle et al.). The percentage of readings is way better and more regular in the log data than in the network data. However, after May 26, this percentage falls in the log data because the logs run out of memory. The network data becomes the best source of information after that date.

#### 2.2.1 Flash Log Data

Let us start by looking at the outliers in the log data. Plotting the humidity against temperature immediately reveals 3 outliers that correspond to nodes 29, 198 and 65535. There is only one datapoint corresponding to the node 65535. It is not clear at all what it corresponds to. All other nodes have many data points

corresponding to measures at each epoch. Furthermore, the humidity is below -9000 and the temperature above 600. Besides that, the voltage is almost zero so this confirms that the battery was not working. For all these reasons, we immediately remove this point from the dataset. Another outlier which we decide to remove is the node 29. The humidity is constantly -4 and the temperature is constantly -38.4 which does not really make sense compared to the other data points. An explanation could be found be looking at the voltage which is always below 2 for this node. In the paper, Tolle et al. explain that in order to produce correct data, the voltage needs to be between 2.4 and 3. Otherwise, the battery is not working properly. Checking the incident PAR (Hamatop) also shows huge fluctuations between two epochs whichs makes us feel comfortable removing this node from the data. Finally, when looking at the node 198 in more detail we see that all the data points look reasonable except a single one where the humidity is recorded -5145.1 and the temperature is recorded as about -8.3. This is only at epoch 3472 which corresponds to May 9th at 6:25 pm. By investigating the battery's voltage, we see that at this epoch, the voltage is 2.33 and this is the only epoch for this node whose voltage is out of the normal range. Therefore, there is no apparent reason why we should remove the entire node. Instead we only remove the datapoint corresponding to epoch 3472.

We then check the voltage. As we said above, it should be between 2.4 and 3 in order for the battery to work correctly. We now investigate nodes whose voltage is beyond this range. Of course we found the 3 outliers we just discussed but there are a bunch of others. After checking the other readings from these other nodes, we could not find anything particularly alarming. For instance, in nodes 128, 134, 135, 141, 142, 143 and 145 the voltage is constant and about 0.58. However, the other variables such as humidity, temperature and hamatop seem reasonable and not completely different from the rest of the data. For these reasons, we choose to keep these nodes and simply set the voltage values to NA.

### 2.2.2 Network Data

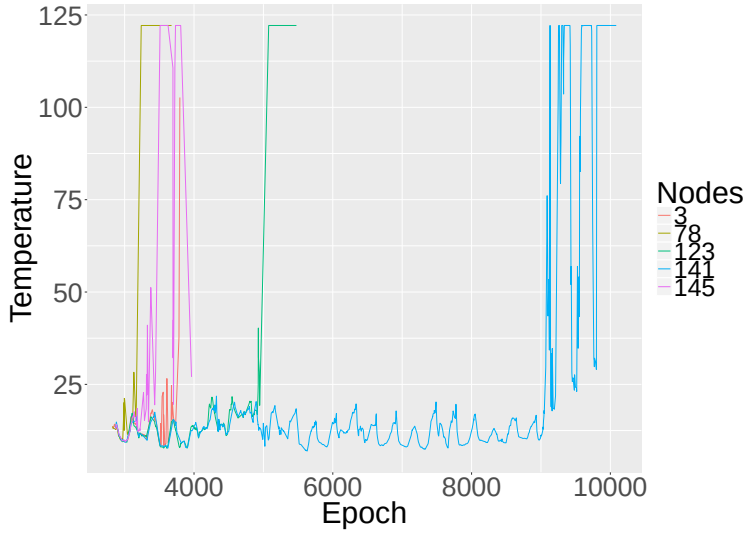


Figure 1: Recorded temperature over time for the suspect nodes 3, 78, 123, 141 and 145 in the network data.

We now look at the network data. While the maximum observed temperature in the log data is 32.58 Celsius degrees, we observe many data points in the network data for which the temperature is above 100 Celsius degrees which does not make sense for the trees. Looking at the voltage here does not really help since they are all very high compared to the log data. It seems like they are on a different scale. The conversion is discussed in the next subsection. However, we note that the nodes that present a strange behavior in the temperature are 3, 78, 123, 141 and 145. We investigate the evolution of the temperature over epochs for these nodes in Figure 1. We notice that nodes 123 and 141 show a reasonable behavior until a certain epoch before it blows up. We then decide to keep data points related to this node up to this epoch. After removal of the other datapoints, the other readings (humidity and PAR) for these two nodes seem reasonable. However the temperature of the three other nodes becomes erratic very quickly so we decide to remove them entirely

from the dataset. Although we do not know yet what the voltage means in this dataset, we still notice that the node 141 has a flat voltage of 1023 which seems very far from all the other data points. We then set the voltage values to NA. The same observation is made about nodes 134 and 135.

### 2.2.3 Merging the two Datasets

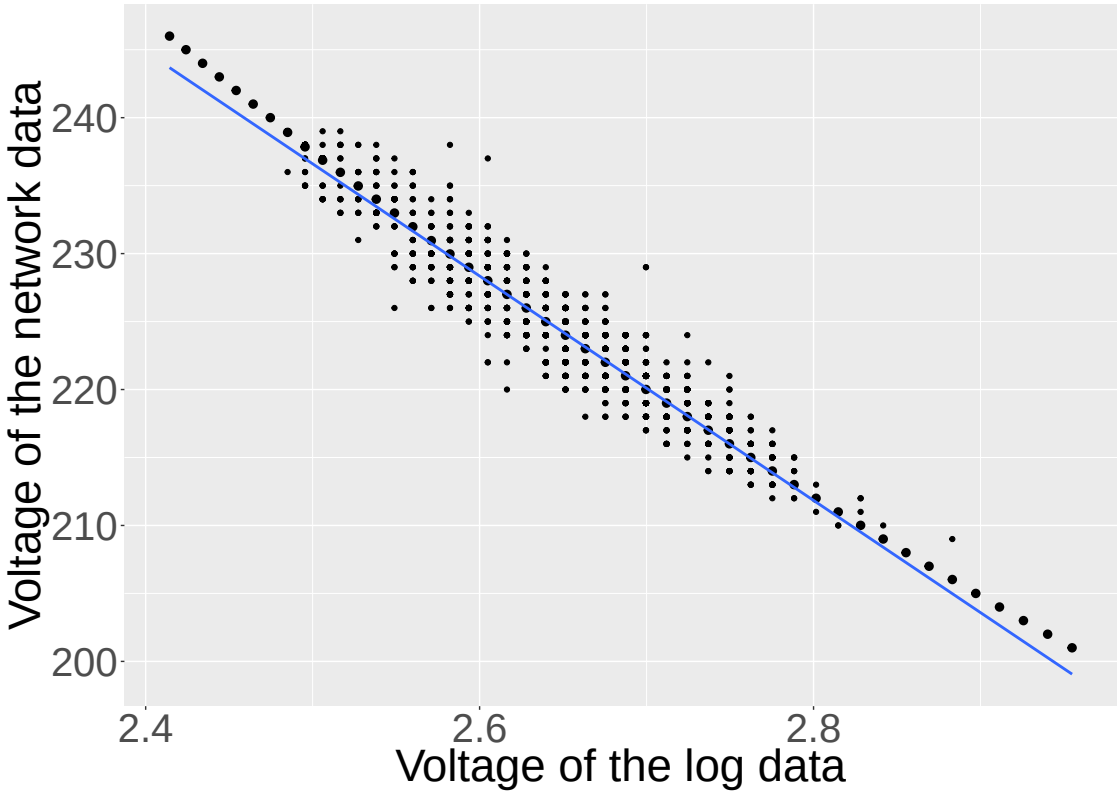


Figure 2: Scatterplot of the recorded voltage in the wireless network data against the voltage in the flash log data. The blue line is the least squares fit.

One of the first things we immediately notice when looking at the network and flash log data is that the voltage looks totally different. In the log data, the voltage seems to be expressed in volts and most of the data points indeed have a voltage between 2.4 and 3, which is the normal range according to the article. On the other hand, most of the voltage values in the network data seem to be around 200. A first guess would be to say that they might be expressed in centiVolts and that we only have to divide them by 100 in order to convert them into volts. However, a simple plot of the net data voltage against the log data voltage shows that this is not the case at all. In fact, the two are negatively correlated ($\rho = -0.9981$) and Figure 2 (in addition to the correlation) shows that the relationship between the two seems to be perfectly linear. For these two reasons, we decide to fit a least squares estimator to the voltage of the flash log data in order to convert the network voltage into the same scale.

The location data frame is also very useful. It provides us with some information regarding the height at which the motes where placed as well as their distance from the trunk and the direction in which they were oriented. It helps us clean a little bit more the data since we remove all the nodes for which the distance from the trunk is bigger than 1 meter. The reason for this is discussed in Tolle et al.: if we consider the sensors that were located more than 1 meter away from the trunk, we might capture broader climate characteristics and not the microclimatic trends that affected the tree directly. Finally, there is one very interesting column in the location data frame called "Tree" that is taking two values: "edge" and "interior". In the article, they are only mentioning one tree but the Figure 1 in the paper presents the interior tree that is 67m tall. We

conclude here that there seems to be a second tree on which sensors were placed and for which data was collected. We may infer that the interior tree is located at the middle of the forest while the edge tree is at the edge of the forest.

Before merging the datasets, we should make sure about a few things. One of them is that we need to avoid duplicates. First we remove all the duplicates in each of the log and net datasets separately, that is the lines which have the same node ID and the same epoch but the most important thing is to remove the duplicates once we merged the dataset as we will have many nodes and epochs for which we have both the network data and the flash log data.

Futhermore, we need to make sure that the dates and times match together. A first thing we notice is that in the log data, the dates and times are wrong. The column corresponding to it always reads "2004-11-10 14:25:00". Fortunately we have a dates data frame that matches epochs with the correct dates. However, there is a second element that is important to notice. For the same epochs, the time reported in the network data is always 7 hours ahead of the time reported in the dates data frame. This is because the network data's dates are according to the Greenwich Mean Time (GMT) while the dates in the dates data frame are in the Pacific Standard Time (PST) zone. We believe that in this case it is more sensible to use times in PST rather than GMT since the trees are in California. As a result, we join the network and flash log data by epoch and match them with the corresponding dates and time using the dates data.

## 2.3 Data Exploration

In this subsection we explore the data more thoroughly.

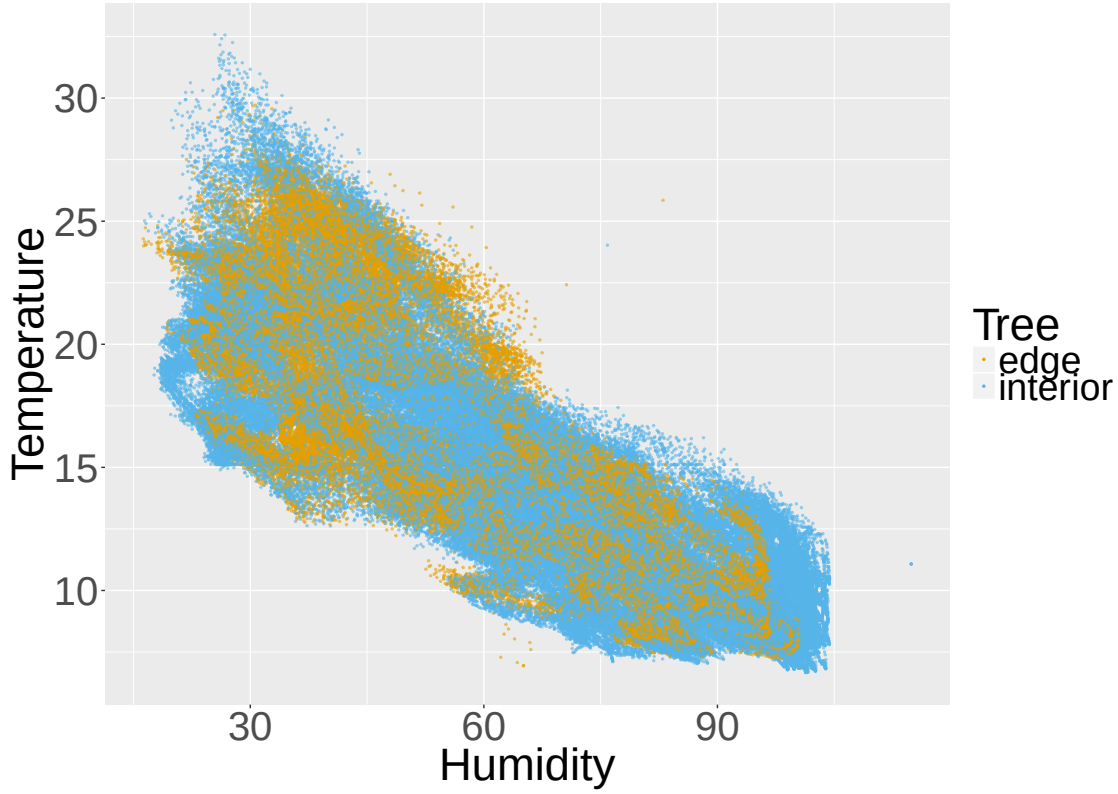


Figure 3: Humidity against temperature. Blue dots are for the interior tree and yellow dots for the edge tree.

We first plot on Figure 6 humidity against temperature and observe that the two are negatively correlated ($\rho = -0.834$). This is actually expected: when temperature rises, the humidity falls. Especially in this region

where the fog is dense, an increased temperature implies an evaporation of the water in the air and a decrease in the humidity; nothing is surprising here.

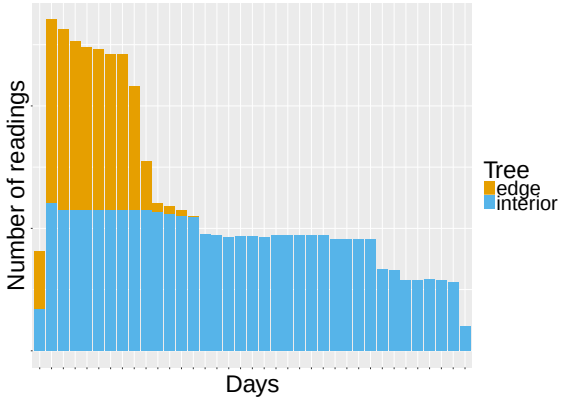


Figure 4: Number of readings per day between April 27 and June 2.

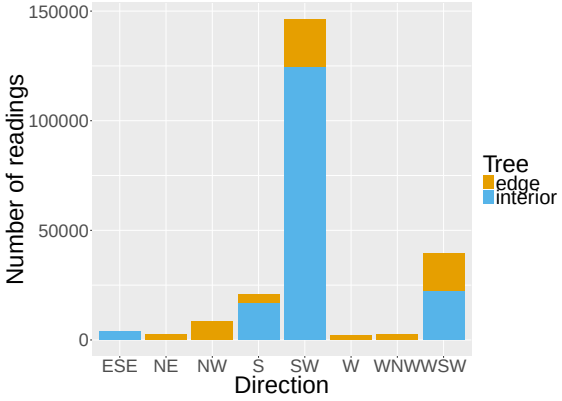


Figure 5: Number of readings for each cardinal direction.

Figure 4 shows the number of readings that are available per days. Counts in blue correspond to the interior tree and counts in yellow to the edge tree. We notice that no measures are available for the edge tree after May 10 while we still have data on the interior tree until June 2. The location data frame provided useful information such as the direction to which the motes were oriented. Unfortunately, data is collected on only half of the compass rose. Furthermore, if we inspect it more carefully, we shall note that most of these directions are on the western part of the rose and the other half is almost not represented. Figure 5 shows that among the 8 directions, only 4 are present in the interior tree data and 7 of them are in the edge tree data. One more observation we should make is that the sensors are not randomly oriented at all. In fact, most of them are oriented towards the South-West.

An interesting plot is the reflected PAR against the incident PAR. First of all, we notice horizontal bands and it is not clear why it is discretized this way. Second, there seems to be a pattern that starts from the left-bottom corner and looks like a parabola and ends at the right-bottom corner. At this point I cannot come up with a good reason why we observe this.

# 3 Graphical Critique

From a strict readability standpoint, I found it very difficult to understand what is going on without reading the article thoroughly. Without talking about the actual content, it took me some time to realize that Figure 3. b) is a list of boxplots. The same blue is used for everything and there are too many small plots. Also the font size is way too small, especially given the number of plots they want to fit in one page. They would better use two or three pages and show bigger plots to improve the readability.

Figure 3 is about trying to describe the distribution of the data through different ways. In Figure 3 (a) they use histograms. Without looking at the visual quality of the graphs, this is rather a good idea. We observe a mode for the temperature around 13 Celsius degrees, a bimodality for the humidity which they explain through the fog that is very common in the area in the early summer. The histograms for incident and reflected PAR show a huge mode at zero which is something we also observed in our exploratory data analysis.

Figure 3 (b) aims to analyze the temporal distribution of the variables by giving the boxplots of each variable everyday. If we use a paper version of the article it is rather hard to see anything interesting but if we zoom in, this is a good way to visualize the temporal distribution for the temperature and for the humidity since they are not too skewed. However, I am not sure if the same plots for the incident and reflected PAR are a good idea. They are not very informative to me. For the reflected PAR, all the boxplots are flat around zero and for the incident PAR, there is not any interesting pattern; the distributions are heavily skewed and the median is constantly around zero.

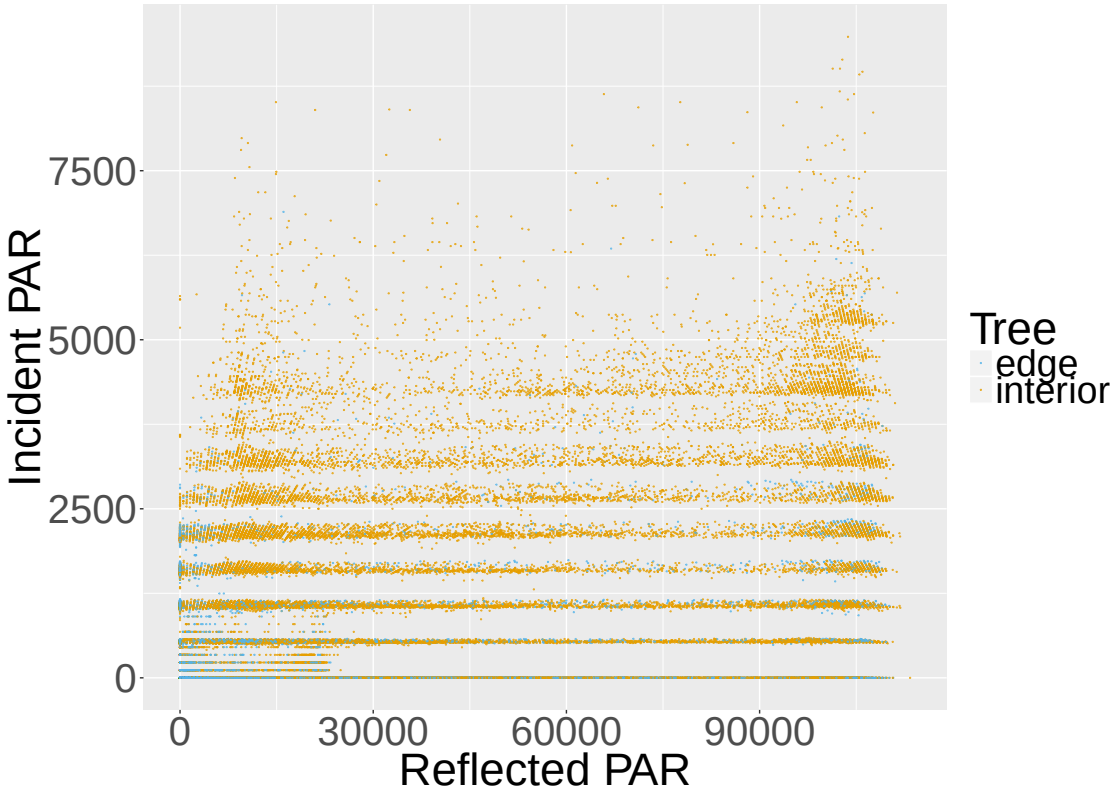


Figure 6: Reflected PAR against Incident PAR.

Figure 3 (c) attempts to look at the spatial distribution of the variables according to the height by giving their boxplots for each height. Here, unlike Figure (b), there is no clear interesting pattern for temperature and humidity and it does not seem obvious that these plots add a real value. Although the boxplots for the incident and reflected PAR are still very skewed, we still notice an increasing trend with the height.

In Figure 3 (d), the researchers sought to correct the issues encountered in Figure 3 (c) by subtracting the timestep mean from the data points. However, I do not think this makes the plot better. I still find the boxplots for the temperature and humidity rather uninformative.

Overall, there are too many plots in Figure 3 and several of them are not necessary. Most of them are exploratory rather than explanatory.

In Figure 4, there are many plots and at first sight, it is difficult to tell what is going on. There is only one caption that reads "Temporal trends and a snapshot in space", which is pretty vague and is not sufficient to understand the plots. Only the first two plots for the temperature and humidity on the left show rather clear trends. I do not think the other two plots for the reflected and incident PAR are very informative. They show that light increases when the sun rises and decreases when the sun sets.

For the right part of the plot, we guess what has been made except that there are no legends at all. In particular, I have no idea what the pink triangles represent compared to the blue ones. Other than that, I think the last plots on the right for the reflected and incident PAR are a good idea to show that they increase with the height although I think it would be more meaningful to do so over several days at the same time. Indeed, here, they consider data points over a single day and of course, there is no sunlight before 6 am or after 9 pm so it does not make sense to consider these points. They simply add many of zero values. Maybe it would be more sensible to choose a specific time of the day where the sunlight is strong and consider different measures of the incident and reflected PAR that day.

# 4 Findings

## 4.1 First finding

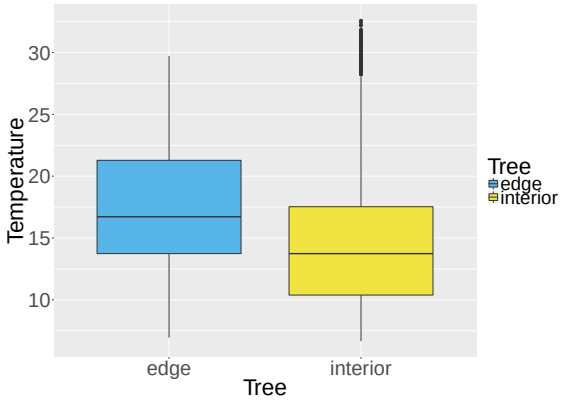


Figure 7: Boxplots of the temperature for the two trees.


Figure 8: Boxplots of the humidity for the two trees.

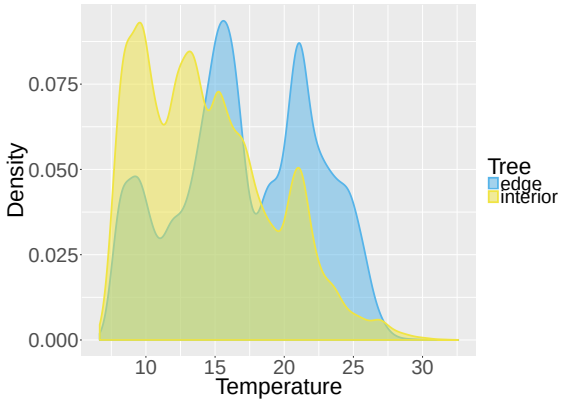


Figure 9: Density plots of the temperature for the two trees.

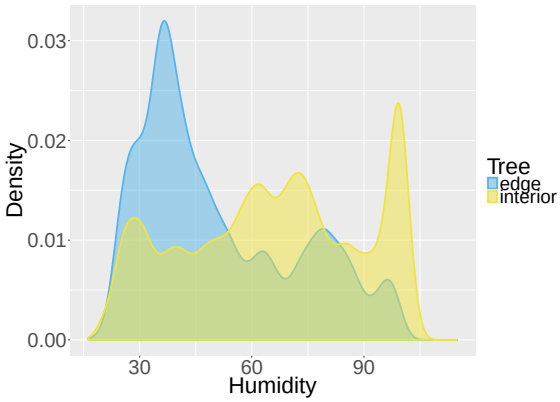


Figure 10: Density plots of the humidity for the two trees.

A first thing we can note is that the temperature is higher on the tree at the edge of the forest that the one at the interior (Figures 7 and 9). This makes sense since it is easy to imagine that a tree at the edge of the forest receives more sunlight than one at the middle of many other trees. We therefore look at the incident and reflected photosynthetically active radiation (PAR) data, hamatop and hamabot but surprisingly this is not the case. It is difficult to tell if this is due to the lack of accuracy of these data – for most of the observations, hamabot and hamatop are zero which is a little surprising. However, looking at the boxplots of humidity (Figure 8) as well as their density plots (Figure 10) confirms our first thought. The humidity is higher for the interior tree than for the tree at the edge. This also makes sense since at the middle of the forest, not only is there more shade but also the foliage of nearby trees tends to keep the humidity much more than the tree at the edge. Furthermore we have already seen in Figure 6 that temperature and humidity are negatively correlated.

## 4.2 Second finding

A second thing that would be expected is that photosynthetically active radiation (PAR) increases with the height. Indeed, the higher the mote, the closer to the sun and the less hidden by foliage. However, plotting all the incident or reflected PAR against the height is not very helpful since we end up with two many data

points that are from every time of the day which makes it very noisy so instead of that, we choose to look at incident and reflected PAR at noon.

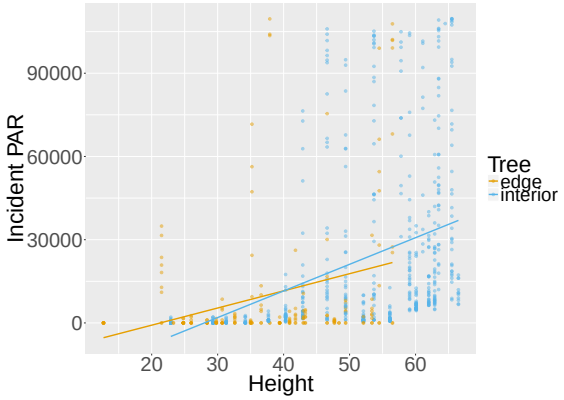


Figure 11: Scatterplot of the edge tree's temperatures for each direction. The blue (respectively yellow) line corresponds to a least squares fit for the interior (respectively edge) tree.

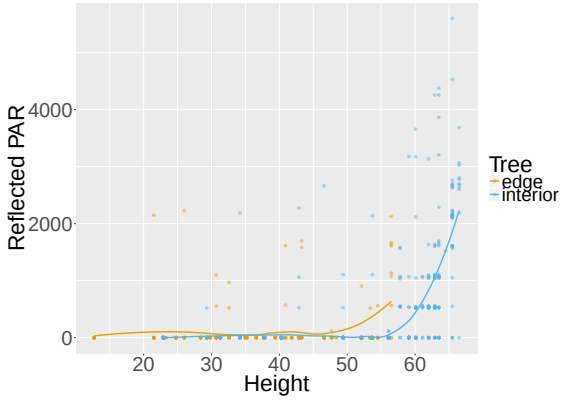


Figure 12: Scatterplot of the edge tree's humidities for each direction. The blue (respectively yellow) curve corresponds to a LOESS fit for the interior (respectively edge) tree.

We observe on Figures 11 and 12 an increasing trend. Also, it is noticeable from the graphs that both the incidence and reflected PAR icrease with the height faster for the interior tree than the edge tree. This might be explained by the fact that the edge tree is more likely to capture light at every height since there are no trees around it to hide the sunlight. However, the interior tree is surrounded by many other trees and only the highest motes are directly exposed to the sunlight.

### 4.3 Third finding

From the boxplots in Figures 13, 14, 15 and 16, we could infer that the tree which is at the edge of the forest might be located at the South-East of the forest: the boxplots show a lower temperature and a higher humidity in the North-West direction. On the other hand, the same boxplots for the interior tree are much better aligned which implies way less variation of temperature and humidity with respect to the different directions of the sensors. While it is rather hard to accurately determine the orientation of the edge tree, especially since we don't have sensors in all the possible directions, the fact that all the boxplots are much more similar for the interior tree than the edge is expected.

This conclusion should be taken with a grain of salt though. Like we said, we would need to have data on sensors oriented in all the directions if we wanted to be more accurate. If we look at Figure 5, we see that only half of the cardinal directions are represented here and we do not have data on the other half. Besides that, only 7 of them are present in the data on the edge tree and only 4 of them are for the interior tree.

## 5 Discussion

This technology is certainly very promising as it enables scientists to collect wirelessly and locally a large amount of data that used to be very hard to gather. However, we have discussed a few issues we had with this dataset. Except for the data cleaning that was particularly tedious for this kind of data, it would have been interesting to be able to compare different trees whose locations were given. We had data on two of them but data on the edge tree is only available on a short period of time (April 27 to May 10) after which we only have information on the interior tree.

Furthermore, the sensors were not evenly placed on the tree since we noticed that the large majority of them were facing the South-West. We do not really have any explanations about this but that would be interesting to know why.

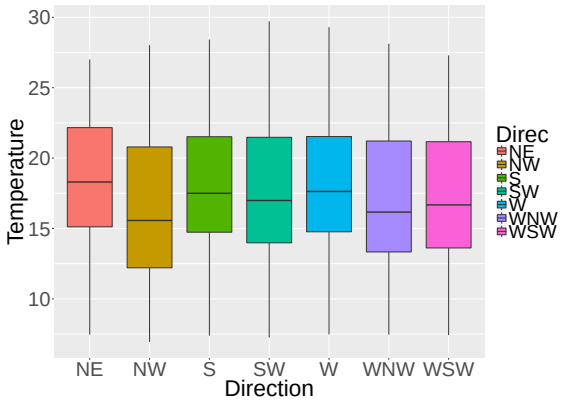


Figure 13: Boxplots of the edge tree's temperatures for each direction.

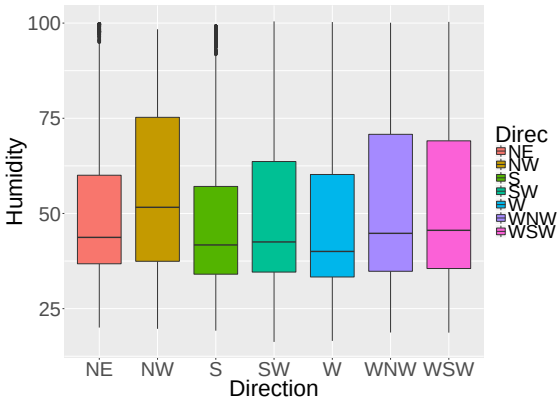


Figure 14: Boxplots of the edge tree's humidities for each direction.

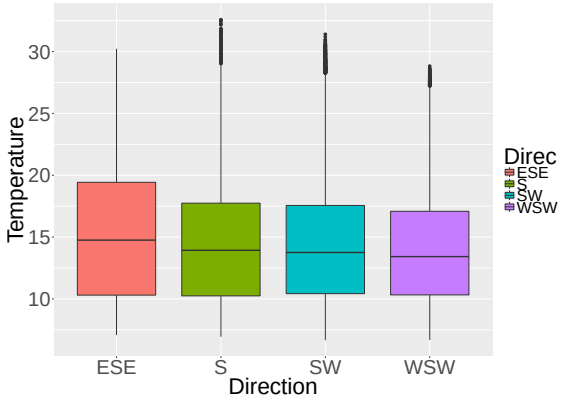


Figure 15: Boxplots of the interior tree's temperatures for each direction.

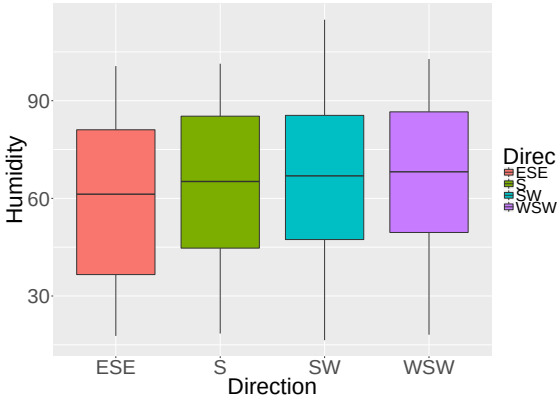


Figure 16: Boxplots of the edge tree's humidities for each direction.

Also it seems like the reflected and incident PAR (Hamatop and Hamabot) often fail to be captured and this results in very sparse data. This is probably due to the fact that the technology is still too recent and improvements need to be made.

# 6 Conclusion

If this technological advance promises a way to better understand microclimates, the exact question formulated by the researchers who wrote the paper remains unclear. As statisticians we would like a clear question of interest to be formulated in order to be able to translate that into a statistical parameter and then try to come up with an optimal way to estimate it from the data that are available. Here this is not the case. We are only given a dataset which we cleaned and investigated. Globally, most results we found were expected.

In future experiments it would be interesting to randomize the position of the sensors on the tree in order to have a more global overview of the characteristics of the trees, so that we know where the temperature is higher, where the humidity is higher, what part of the tree best captures the sunlight.

# Informal Collaborators

Some discussions with David Chen, Yutong Wang and Nicholas Sim.

# References

Tolle, Polastre Et Al, "A Macroscope in the Redwoods"