Lab 1 - Redwood Data
Stat 215A, Fall 2017

SID: 25928461

1 Introduction

This report will examine the macro-climate of two Redwood trees located in Sonoma, California over approximately one month. The data was collected by Tolle *et al.* [1] in 2004 and unfortunately contains a large number of data entry errors and inconsistencies. As a result, a large amount of data cleaning was performed prior to analysis. In Section 2, the data as well as the data cleaning procedures are described. In Section 3, the figures presented in the original paper by Tolle *et al.* are discussed in terms of what questions they were trying to answer, how successful they were in answering them and whether or not they raised any questions not addressed in the original text. The main findings arising from the data anlaysis are described in Section 4, and each finding is accompanied by a descriptive graphic.

2 The Data

The raw data contains a large number of measurements collected from two trees in the Sonoma region (one tree was located at the edge of the forest, and the other in the forest interior). Data was collected both from a wireless sensor network (the network dataset) and from a data logger (the log dataset). The data logger supplied data from 39 nodes located on the tree at the edge of the forest and 30 nodes located on the tree in the interior of the forest. The wireless sensor network, on the other hand, provided measurements from 29 nodes located on the interior forest tree but no measurements were recorded for the tree on the edge of the forest.

2.1 Data Collection

The data collection procedure is described in detail in Tolle *et al.*, and is summarized here. Each tree contained a network of nodes deployed at various locations throughout the tree. Each node in the network contained a number of sensors which took measurements of air temperature (^{o}C), relative humidity (%) and PAR (photosynthetically active solar radiation) from both on top of the node (incident PAR) and below the node (reflective PAR). This report will focus primarily on temperature, humidity and incident PAR. The nodes were located at various heights (ranging from 15m above ground level to 70m above ground level, with roughly a 2m spacing between nodes), various angular locations (although most of the nodes were facing South, South-West and West-South-West) and various radial distances from the trunk (most were very close to the trunk, however a small number were placed further out). Throughout the study, each sensor was designed to take a measurement every 5 minutes, for which the network was awake for 4 seconds while the data was collected and transferred back to the base station.
Each node contained a battery and two sensor boards confined within a sealed cylindrical enclosure designed to reflect most of the radiated heat. One sensor board captured radiation and the other captured temperature and humidity. The sensors were calibrated and subsequently reset prior to deployment. The nodes began recording prior to installation to ensure that the nodes were operating correctly and to ensure that they could all synchronize with the gateway receiving the data. However, as we will see below, there are large parts of the study where either a number of the sensors were not working, or the data collected was not recorded.

2.2 Data Cleaning

Unfortunately there were a number of misleading and erroneous measurements contained in this dataset. For example, there were large numbers of missing values, resulting in entire days with no data recorded. There were also several ambiguous variables as well as a number of measurements which made no sense.
According to Tolle *et al.*, it was intended that the data collected by the data logger and the wireless network were equivalent. However an initial glance at each dataset showed that the data logger recorded significantly more measurements than did the wireless network. In particular, the data collected by the logger contains 301,056 observations while the data collected by the network contains 114,980 observations, which is less than half of that recorded by the data logger.
Next, a brief comparison showed that the data logger dataset contains measurements on a total of 72 nodes (30 from the interior forest tree, 39 from the edge forest tree, and 3 nodes for which it is unclear to which tree they correspond), while the network dataset contains measurements only on 31 nodes (29 of these are located on the interior forest tree, 1 is located on the edge tree but contains only 4 measurements and so is probably an error, and there is 1 node for which it is unclear to which tree the node corresponds). 28 of the nodes were common to both the wireless network and the data logger datasets.

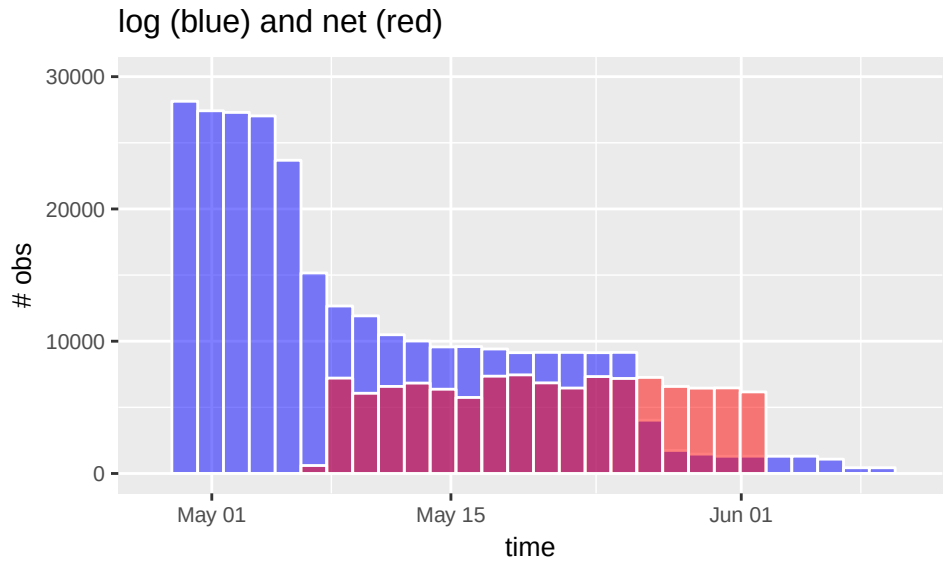

Figure 1: Histogram showing the number of observations for the net dataset (red) obverlaid onto a histogram showing the number of observations for the log dataset (blue) versus time

Of these nodes, there are 10 nodes for which the data logger contained less than 1,000 observations and 8 nodes for which the network provided less than 1,000 observations. Given that the expected number of observations is approximately 12,000, this corresponds to a lot of missing data. The number of overall observations made over time from both the data logger and the wireless network is presented in Figure 1. Unfortunately, there are approximately two weeks (one week at either end of the 44-day study) for which the data logger collected measurements and the wireless network did not. Although the number of observations received via the wireless network remained fairly constant (when data was recorded, that is), this was clearly not the case for the data logs. After May 5th the number of observations made daily approximately halves, and in the days following May 25th, the number of observations made by the log dataset was reduced to almost none. According to Tolle *et al.*, this corresponds to when the data logs "filled up". Overall, for the observations that are common to both the data obtained from the data logger and the wireless network, the measurements are very similar (although the voltage reported in the network dataset has an inverse relationship to the voltage reported in the log dataset such that $\log \text{voltage} = \alpha + \frac{\beta}{\text{net voltage}}$). Thus this report will focus on the data received from the data logs for values prior to May 26, and will not examine the wireless network data in any great detail.

The data contains the following variables:

- `result_time`: the time at which each measurement was taken
- `epoch`: ID number for each measurement time
- `nodeid`: ID number for each node
- `voltage`: voltage
- `humidity`: relative humidity (%)
- `humidity_temp` (renamed to be `temp`): temperature (^{o}C)
- `humidity_adj`: undefined
- `hamatop`: incident PAR
- `hamabot`: reflective PAR

as well as `parent` and `depth`, whose definitions are both cited as "network structure", however no further explanation was provided either in the documentation supplied with the dataset, nor in Tolle *et al.*'s paper. With this in mind, `parent` and `depth` were removed from the data. The variable `humid_adj` appeared to be extremely similar to the `humidity` variable and so was also removed. Strangely, there are a number of `humidity` observations which are either greater than 100 or less than 0. Intuitively this makes no sense as relative humidity is defined percentage and so should be between 0 and 100. However, the humidity values above 100% visually appeared to fit with the data so were left alone and it was assumed that these values were due to a minor calibration error. The humidity values below zero, on the other hand, appeared to correspond to measurements recorded by a faulty node, which displayed strange voltage behaviour and which also recorded temperatures of approximately -40^oC and incident PAR measurements much higher than seemed reasonable (see Figures 2 and 3).

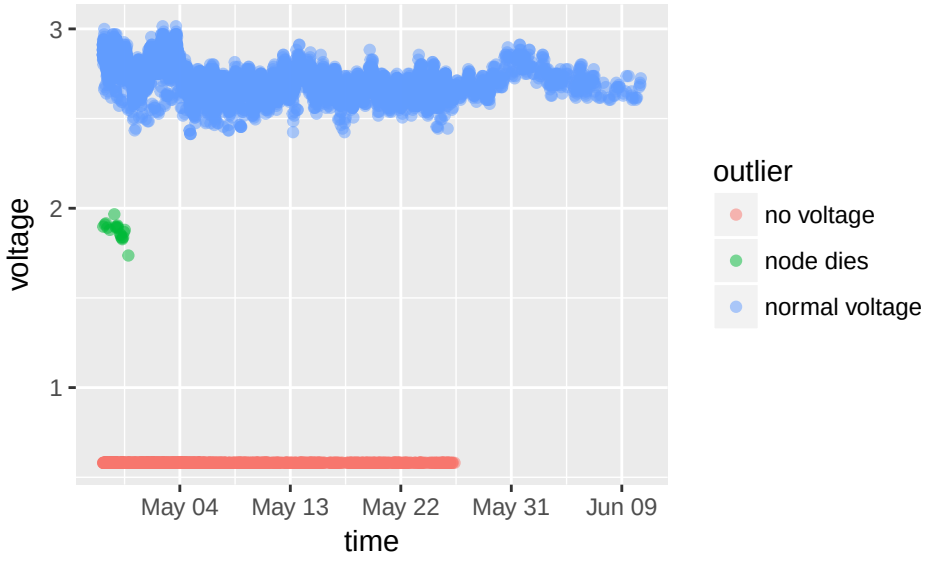

Figure 2: Voltage versus time for the data obtained from the data logger

Although Tolle *et al.* discarded all nodes whose voltage readings dipped below 2.4, for this report, only one such node was discarded, since, as mentioned above, a closer look showed that this node was giving the erroneous humidity and temperature measurements. There were a number of other nodes whose voltage readings were constant at 0.6, however, the other measurements taken by these nodes appear to be within

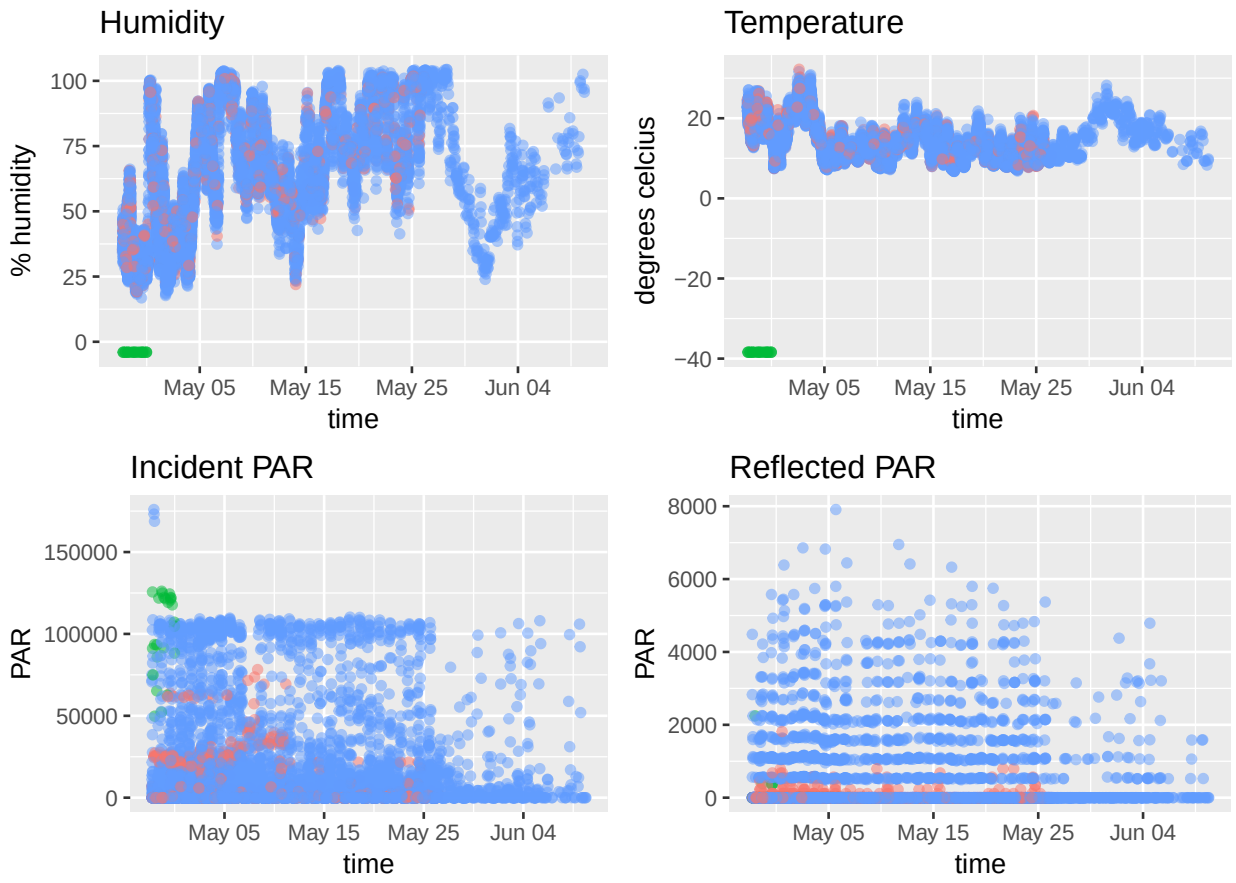

Figure 3: Scatterplot of each variable versus time in the data logger dataset. The color of each point corresponds to Figure 2

normal ranges, and so no action was taken (Figures 2 and 3). A summary of the nodes which were entirely removed is presented in Table 1.

Since the study was designed such that one measurement was taken every 5 minutes, one would expect to see approximately 12,000 time stamps. In the log dataset, however, the `result_time` variable contained only one date (14:25:00, November 10th, 2004), repeated for each observation. Fortunately, an external file containing the correct `epoch` and `result_time` combinations was available and was incorporated into each dataset to reflect the correct measurement times.

Next, the dataset contains a large amount of missing values. For the most part, the missingness comes from the fact that there are a huge number of node/time combinations missing from the dataset, but also that there are 8270 observations over three nodes in the log dataset with `NA` values reported for each of the measurements taken. Obviously the simplest thing to do in such a situation is to simply remove the missing value rows. However since more than 60% of the observations made by these three nodes were `NA`, all observations from each of these nodes were removed. The reason being that having such a large amount of missing data implies that the node was not functioning correctly for a large portion of the study, and that the rest of the observations are potentially unreliable.

2.3 Data Exploration

There are a number of differences between the two trees, the first being that they are slightly different heights. The interior tree contains nodes at heights ranging from 22.9 meters to 66.4 meters above ground

	nodeid	n	NA	humid	temp	hamatop	voltage
1	2	65	0	0	0	0	0
2	15	2344	68	0	0	0	0
3	24	57	0	0	0	0	0
4	27	81	0	0	0	0	0
5	29	660	0	100	100	47	100
6	40	98	0	0	0	60	0
7	62	73	0	0	0	0	0
8	109	345	0	0	0	0	0
9	114	536	0	0	0	0	0
10	122	8109	65	0	0	0	0
11	128	2204	62	0	0	0	100
12	200	65	0	0	0	0	0
13	65535	1	0	100	0	0	100

Table 1: The first column provides the node IDs of each node which is removed from the log dataset. The remaining columns display the number of observations, the proportion of NA values in the dataset, the proportion of erroneous or outlying humidity, temperature, hamatop and voltage readings respectively for each node

level, whereas the edge tree contains nodes at heights ranging from 12.7 meters to 56.5 meters above ground level. Moreover, the bulk of the observations within the log dataset come from the interior tree and the edge tree only contains data for the first week of the study (Figure 4).

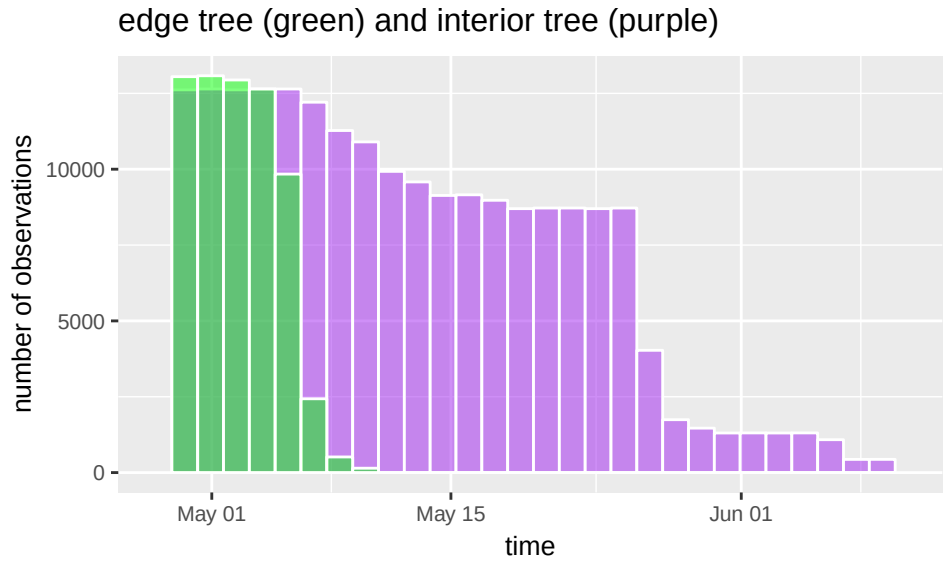

Figure 4: A Histogram showing the number of observations in the log dataset for the tree in the edge (green) of the forrest overlaid onto a histogram showing the number of observations for the tree at the interior (purple) of the forrest versus time

To get an initial impression of how the variables interact, Figure 5. presents a scatterplot of each of the variables from the interior tree. The corresponding scatterplot from the edge tree was extremely similar. There is a clear overall inverse relationship between humidity and temperature such that as humidity rises, temperature decreases (temperature and humidity have a correlation of -0.84). A closer look at the data (see Figure 7 for an example) revealed that the troughs and peaks occurring in the temperature measurements precede the corresponding peaks and troughs occurring in the humidity measurements by approximately 2 hours. It is harder to specify a strict relationship between incident PAR and either humidity or temperature,

other than lower values and less variation in the incident PAR values occured when temperature is low and humidity is high. As we will see below, this corresponds to the night hours, when there is typically no incident PAR detected (as there is no sunlight) and the temperature drops. The large majority of the incident PAR measurements were very low.

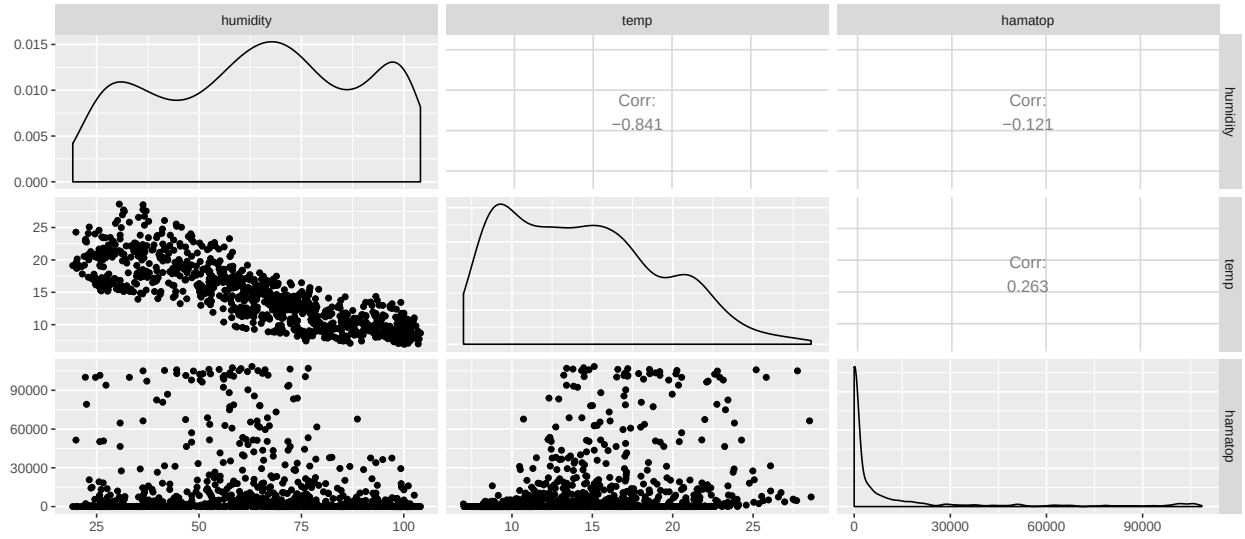

Figure 5: Paired scatterplots for a subset of the temperature, humidity and incident PAR (hamatop) measurements from the interior tree

3 Graphical Critique

Tolle *et al.* view the data as consisting of three primary dimensions: time × height × value. However, the authors acknowledge that visually, it is not an easy task to make inferences by looking at everything at once. Thus, quite sensibly, they decide to split the dimensions in several ways. The first is by looking at the one-dimensional value (e.g. temperature, humidity, PAR) alone via histograms. The next is to use boxplots to analyse the two-dimensional time × value and height × value combinations in order to examine how the measurements change over height and time separately.

In Tolle *et al.*'s Figure 3(a), the authors aim to identify what type of distribution each variable follows. In terms of temperature and humidity, the histograms offer clear visualizations of the overall distributions. However for the PAR histograms, the figures are so dominated by the vast percentage of values at 0 that it is impossible to get a clear picture of the remaining values. The authors may have been better served by considering a log-scale for the PAR values obtained.

Overall, these histograms serve only to get a feel for the range of values attained for each variable, which could be presented just as clearly in Tolle *et al.*'s Figure 3(b) which displays the variation in each measurement over time in the form of boxplots. As intended, for temperature and humidity, these figures display the range and change in values observed over time. However, it would be much clearer if a line of best fit were added to the plot making the overall trends visually clearer. Further, unlike in Figure 3 in this report, these histograms do not give an idea of how many observations are made on each day. In particular, it is not clear that after May 26th, the number of observations decreases significantly. Again the PAR figures present much of the same information as they did in Tolle *et al.*'s Figure 3 (a).

For Tolle *et al.*'s Figure 3(c), the boxplots are stacked vertically, presumably to visually represent tree height. Although these boxplots do provide a visualization of the changes for each variable over height, they are somewhat cluttered and do not allow for eyeballing of any particular trend. Especially as these values are taken from the entire study rather than for any particular day, most of the interesting trends get lost since each day is so variable. Comparing Tolle *et al.*'s Figure 3(c) to Figure 8 in this report, it is much easier to see any interesting trends in Figure 8 as we are looking at directly comparable measurements rather than

"averaging out" over the entire study.
Figure 3(d) in Tolle *et al.* attempts to show variability after centering (by removing the effect of the mean). Unlike in the previous panels, here the temperature and humidity plots are visually uninformative, but for the PAR panels, the relationship between PAR and height is presented in a much clearer manner. From this plot it is fairly clear that the lower segment of the tree receives much less light than does the upper segment of the tree. Perhaps the authors could have removed part (a) and combined part (c) and (d) by taking the temperature and humidity panels from part (c) and the PAR panels from part (d) in order to present the same information in a more concise manner.
Figure 4 in Tolle *et al.*, on the other hand, contrasts with Figure 3 in Tolle *et al.* in that it considers changes in only one day in the study (May 1st). Unlike in Figure 3 where it was difficult to pick out any clear relationship between the variables, this figure makes it much easier to identify how the variables interact with one another. The left-hand panels represent each individual sensor as a colored line, which while it looks somewhat messy, does allow for direct comparison between temperature and humidity, for example as the red sensor reaches a temperature peak, it also reaches a humidity trough. However the majority of the other sensors are lost in the jumble of colored lines.
The panels on the right-hand side of Figure 4 combine height, value and direction by selecting the points to be triangles whose direction represents the direction of the node. However, since this figure considers only a single time-point, it is unclear as to whether the trend presented is an accurate description of all of the data or simply something that occurred by chance. In other words, the figure does not capture any of the variability of the trends presented, and may present an extremely biased view of the relationship between height and the other variables. These figures raise the question of whether these trends presented are representative in general to the tree or only representative to May 1st at 9:35 AM. Unfortunately, this question is not addressed in the paper.

4 Findings

Below we will describe three of the primary findings made primarily from graphical analysis of the dataset. The first finding presented is the variation in the relationship between temperature and humidity over time. The second finding describes an anomaly in the data occurring in late April for the forest edge tree which indicates a possible fog cloud that descended onto the forest edge, but left the forest interior unaffected. The final finding describes how the temperature, humidity and incident PAR varies over different heights of the tree.

4.1 Temperature versus Humidity

Although Figure 5. presented an overall relationship between temperature and humidity, the story is a bit different if we take into account the time of day. Figure 6. displays humidity versus temperature over four consecutive days and displays the time of day through coloring (dark points correspond to nighttime and light points correspond to daytime). During daylight hours, the relationship between temperature and humidity was significantly more linear than during the night hours. It appears to be the daylight hours which drive the overall trend observed in Figure 5. During the night hours, each plot contains two distinct temperature-humidity clusters, corresponding to early morning and late night (note that it is expected that these two clusters would arise since they are separated by the daylight hours). Although the night clusters still often display this inverse relationship between temperature and humidity, this trend is not nearly as steep nor as linear as during the day. For example, although the temperature decreases as humidity increases during the day on May 2nd, at night, the temperature remains fairly constant around 20^oC even with changes in humidity. It is clear that the overall trend is for the night to be cooler and less humid than the daytime.

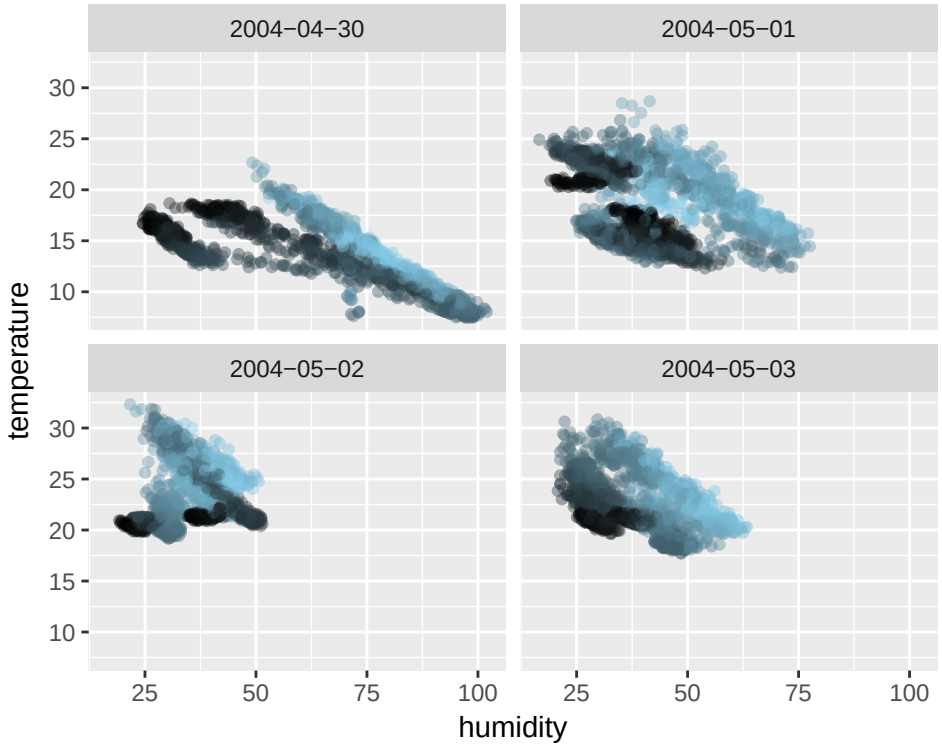

Figure 6: Temperature versus humidity for the interior tree on each of April 30th, May 1st, 2nd and 3rd. The points are colored based on the time of day, and intended to reflect the color of the sky. A black point corresponds to midnight, while a sky blue point corresponds to noon. There is a continuous color scale for all times in-between.

4.2 Late April Fog

Comparing the edge and interior trees on April 28th-29th, one would get the impression that the two trees experience wildly different conditions (first column of Figure 7). April 28th begins with a steep drop in temperature in the environment surrounding the edge tree, accompanied by a corresponding increase in humidity. Meanwhile, the interior tree experiences the inverse: a slight increase in temperature and a decrease in humidity. Around 7AM the temperature around the interior tree begins to decrease, reaching a minimum at 10AM after which the temperature rises to a maximum at 4PM. Following this, the temperature gradually decreases as night falls. Very different behavior is observed for the edge tree during this time. In between 8AM and 1PM, the temperature around the edge tree sharply increases after which it decreases to a local minimum at 4PM and increases again to the maximum at 7PM before decreasing steadily as the turbulence of April 28th comes to a close. Having described the temperature variations over the course of the day, one can infer the humidity variations, or better yet, one can look at Figure 7 and see for themselves the same (though inverted) patterns made for the edge and interior tree.

As interesting as these vastly different humidity and temperature observations are, viewing the incident PAR values over those two days reveals that the sensors from the edge tree are detecting PAR levels of zero throughout the entire day, whereas the sensors from the interior tree are detecting regular PAR levels. Although it is certainly possible that these readings are simply the sensors misbehaving, one possible explanation for this observed phenomenon is that a dense fog descended onto the edge of the forest where the edge tree lives (resulting in the PAR levels detected being negligible and the increase in humidity) while the interior tree was deep enough inside the forest to be unaffected by the fog.

Although this section has discussed in length the strange weather behavior of April 28th and 29th, it turns out that on all other days for which data is available from both trees, the two trees experienced strikingly similar temperature, humidity and PAR trends to one another (although it can vary significantly from

Figure 7: Temperature, humidity and incident PAR (hamatop) over time on April 28th for which the two trees appear to experience very different environments and on May 4th when the two trees experience a very similar environment.

day to day). The edge tree, however, consistently appears to experience a slightly cooler and more humid environment during the day than does the interior tree (see the second column of Figure 7).

4.3 The Effect of Height

Figure 8 shows how temperature, humidity and incident PAR varied for three different height regions of the interior forest tree on April 30th and two days later on May 2nd. Interestingly, on the cooler day (April 30th), the the upper region of the tree experienced temperatures of a few degrees warmer and slightly lower humidity than the lower region of the tree, while on May 2nd, which was significantly warmer, there was little difference between the temperature and humidity of the upper, middle and lower regions of the tree. Further, the lower and middle regions of the tree were exposed to slightly more sunlight on the cooler day, while the upper region of the tree was exposed to more sunlight on the warmer day. As expected, on both days the amount of sunlight detected by the sensors increased significantly the higher up the tree the sensor was located.

5 Discussion

Although the dataset provided was fairly large, it was not so large that visualization was made impossible. Rather, there were a number of logistical factors that made analysis challenging for this particular data set. For example, there was a lot of ambiguity about what several of the variables represented (for example, it was not clear that the tree variable represented two different trees). There was also a large amount of missing values and erroneous data entries, which were not necessarily obvious from looking only at summary statistics. Further, the two data sources (the network and the data logger) were inconsistent with one another, for example, in terms of the measurements units used.
Overall, it seems that in terms of large datasets, it is much more useful to use informative graphics to understand and explore the dataset than it is to look at numbers such as summary statistics.

6 Conclusion

In this report, an analysis of the climate of two redwood trees from Sonoma, California was performed. The data contained a large amount of inconsistencies, incorrect entries, missing values and outliers and the corresponding data cleaning process was described in detail. Following data cleaning, visual analysis lead to several interesting findings including 1) that the relationship between temperature and humidity depended on the time of day, 2) that strange conditions in late April potentially indicates heavy fog surrounding the edge of the forest, and 3) that it is slightly cooler and more humid around the lower region of the tree than the upper region of the tree, but this difference is more extreme on cooler days.

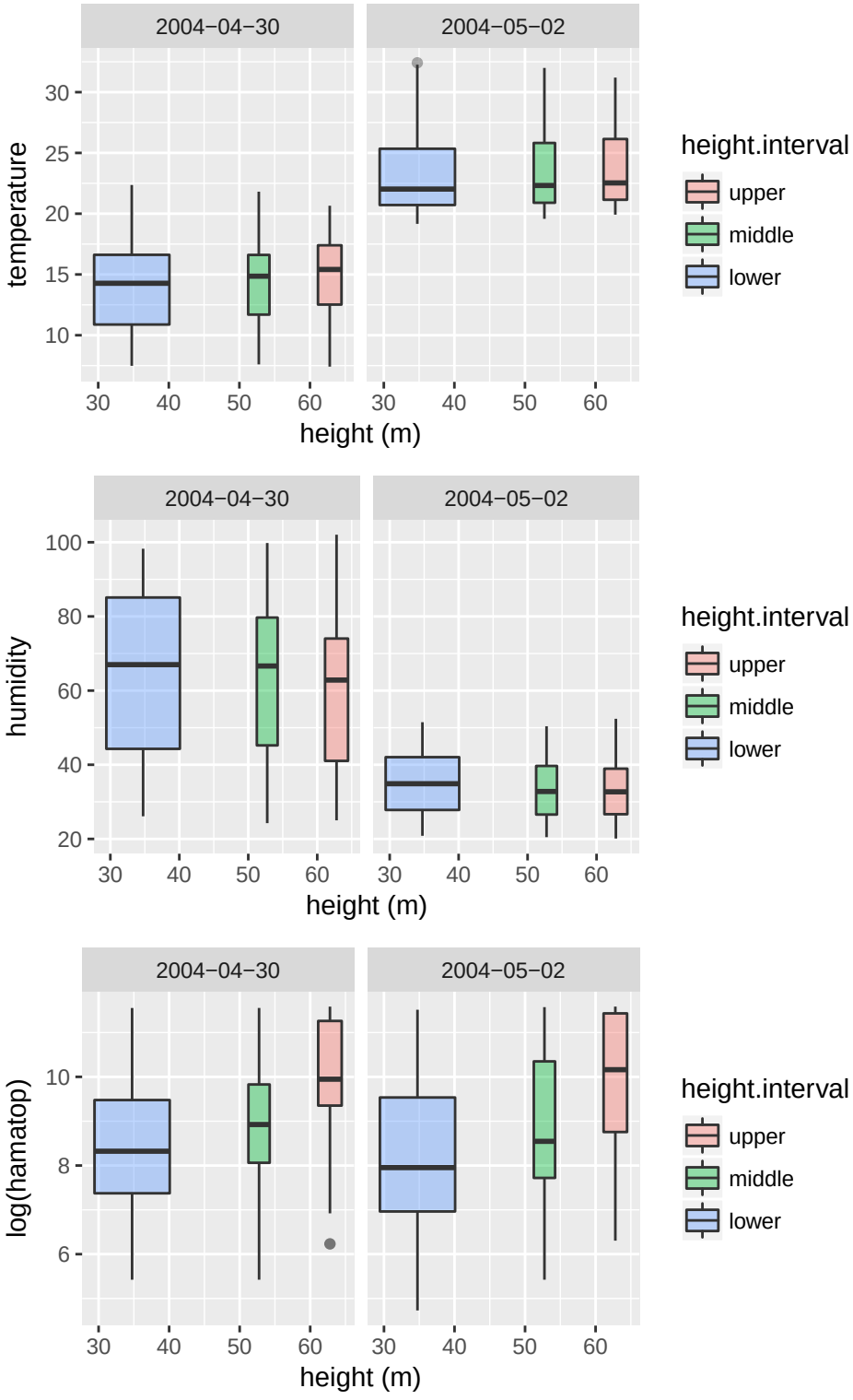

Figure 8: Boxplots displaying the temperature, humidity and log(incident PAR) on April 30th and May 2nd for the interior tree at the lower, middle and upper regions of the tree

References

[1] Tolle G, Polastre J, Szewczyk R, Culler D, Turner N, Tu K, Burgess S, Dawson T, Buonadonna P, Gay D, Hong W: **A Macroscope in the Redwoods**. In *In Proceedings of the 3rd ACM Conference on Embedded Networked Sensor Systems (SenSys)*, ACM Press 2005:51–63.