Redwood Data Analysis, Stat 215A, Fall 2018

Aummul Baneen Manasawala

September 28, 2018

1 Introduction

This report is about reproducibility and analysis of the "A Macroscope in the Redwoods" paper[6]. We also include three interesting findings from the data collected by the authors in the case study of a wireless sensor network that recorded 44 days in the life of a 70-meter tall redwood tree, at a density of 5 minute in time and every 2 meters in space. The sensor network is referred as 'macroscope' as it can perform dense temporal and spatial monitoring that would help gain insights of the complex interactions. Using this technology, the authors of the paper[6] report microclimatic monitoring of a coastal redwood canopy. In doing their study, they use multidimentional analysis methodology to more deeply understand the dense and wide-ranging spatiotemporal data obtained from the macroscope. In this report we would start with inspecting the data of their study, its collection, cleaning and exploration. It would be followed by the three intereseting findings from the data which are different to the ones mentioned in their publication.

2 Motivation

The authors of the paper[6] chose to study the microclimatic changes through the temporal and spatial densities of the redwood amongst a lot of natural phenomenons that could be studied using the macroscopic network. The reason for the same lie in the fact that the top of the tree experiences a wide variation in the temperature, humidity and ofcourse, light while the bottom is typically cool, moist and shaded. This creates a non uniform gradient in the weather front that move through the structure of the tree. When the sun rises, the top of the canopy warms quickly. This warm front moves down the tree over time until the entire structure stabilizes or until cooling at the canopy surface causes the process to reverse. Humidity front also move through the canopy but the process is less complicated by the tree moving so much water up from the soil and into the air. At some point, the observed humidity is driven by the transpiration process and decoupled from the prevailing climatic conditions.

3 The Data

The data can be roughly divided into two on the basis of the value of ineterst and the deployment characteristic. In the set of the data which is the value of interest is the data that is gathered by the Internet of Things (IoT) sensors. This includes the temperature, humdity, incident PAR (Photosynthetically Actice Radiations) and reflected PAR. These are the climatic parameter's whose distribution over space and time we need to study. The second part of the daat that consists of the deployment characteristics is the daat that accounts for the differences in the treatment and all the confounding factors that might be affecting the values that we get for the climatic parameters of ineterest. This set of data consist of the time at which the reading from the sensor is recorded, height of the sensors from the ground level, angular location of the sensor and the direction they face and the radial distance of the sensors from the tree.

3.1 Data Collection

The author of the paper[6] designed wireless micro-weather station based on the Berkeley Motes manufactured by Crossbow. They stood on the shoulders of giants and made use of the established technology provided by TinyOS[3], MiniRoute[7], and TinyDB[4]. The software for the node operating system, network stack and data collection framework was TinyOS ans TASK.

The data was collected from the sensors once every 5 minutes because they felt it was sufficient to capture the variation. The entire time frame of the data collection was 44 days and this task was carried out by the authors of the paper themself. In order to capture the gradient in enough detail, nodes were placed with roughly 2 meter spacing between them starting from 15m from the ground to 70m from the ground level. Most nodes were on the west side because the west side of the canopy was thicker and provided the most buffer against direct environmental effects. In terms of the radial distance of the nodes,they were pretty close to the trunk and the distance range was 0.1m to 1.0 m only. This was done in order to ensure that the microclimatic trends that affected the tree directly were captured and the effects of the broader surrounding climate could be ignored. This being said, some of the nodes were placed outside of the angular radial envolope to monitor the microclimate in the vicinity of other biological sensing equipment previously installed.

3.2 Data Cleaning

The data, since it is obtained from the Internet Of Things (device), is not very clean and is highly erroneous. There are a lot of missing values and a lot of values which doesnot make any sense. We approached data cleaning with a lot of steps such as removing the missing values, removing values not possible in the real world etc and we would dive further in the intricacies of some of them.

Firstly, we first start with dealing with the missing readings. We had two options to attack the missing values, either to remove the row with a missing value entirely or to replace the missing value with another reading. If we remove all the rows with the missing values, we loose roughly 20 percent of the data which is a setback. But, since we are looking for patterns in the data, if we chose to replace the missing values with a value, say mean, or median, it might negatively interfere in the patterns and skew our inferences. Therefore, we chose to go with the smaller of the two evils and remove the missing data.

Secondly, we found out that the humidity values were negative and also larger than 100 which is not possible in the real physical world for a 70-meter height[5]. We referred to the paper[6] and trusted their range of humidity which starts from 16.4. We remove all the rows of the sensor reading for which humidity is recorded less than 16.4 as clearly it is a faulty reading and there is a high chance of the other values recorded by the sensor at that moment to be erroneous too.

Similarly, we follow the guidelines of the authors of 'A macroscope in the Redwoods' for the temperature which according to them is in the range of 6.6 to 32.6 degree celcius. We remove all the data points outside this range without loosing more than 0.03% of the data.

Coming to the Incident PAR, the readings are too hayward than to the range described in the paper. Authors Tolle et al. suggest that the incident PAR must lie in the range of 0 to 2154 but the data doesnot confirm. After cleaning the data for bashing outliers like negative values and values as larger than 50000, we still have 30% of the data as seen in figure 1 not falling in that range. There must be something seriously off either with the units of measurements reported in the paper and the units of measurement of the data or normalization. We could not confirm to the norms laid out by the paper because the massive extend of non confirmity of the data to those recommendations. We would like the author to revise the range or data to explain this so that we don't loose as much as 36% of the data in our anlysis. For the purpose of our analysis, we choose the range to be from 0 to 10000 to loose only 20% of the data which is also a lot but atleast we are assured of the data not being too wrong which could otherwise adversly affect our analysis and lead to faulty outcomes.

For the reflected PAR, we face the same dilemma. As much as 15% of the data donot fall in the range described by the authors for the reflected PAR i.e. 0 to 180. However, since we have a large dataset and it is feasible to do the analysis even after loosing that, we trust the author's recommendations and remove all the data with the reflected PAR lower than 0 and higher than 180. We would highly want the clarifications

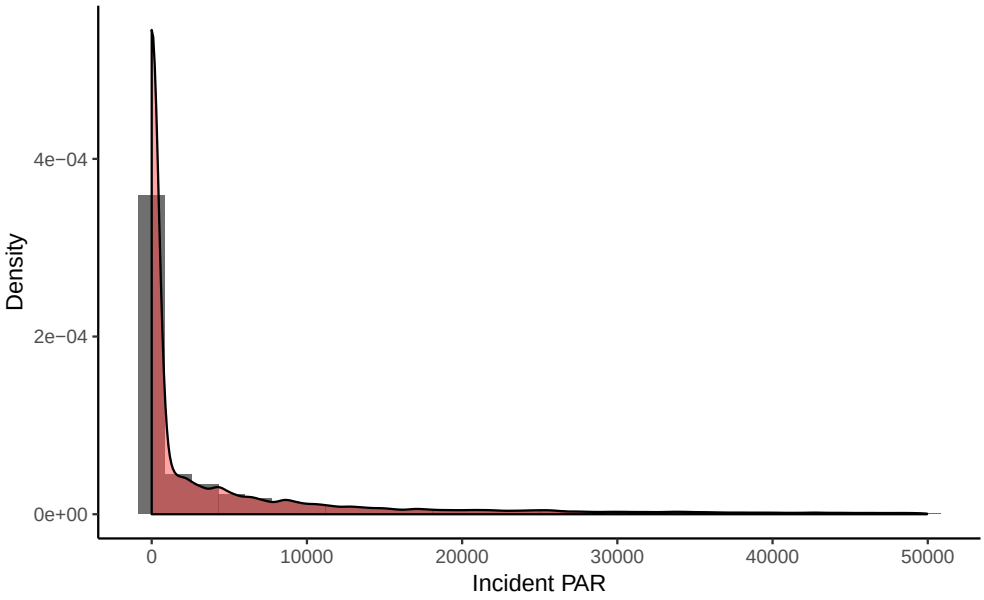

Figure 1: Distribution of Incident PAR

from the author about the heavy percentage of the data deviant from their range, but because of lack of access, we take the safer side and perform the analysis on the cleaning lines that the authors suggest.

Most important of all, the erroneous readings and the outliers were correlated to the voltage of the batteries[6]. In the section 5.6 of the paper[6], authors mention that there is a positive correlation between anomalies and the battery failure. We removed all the voltage readings lower than 1 to ensure that we don't include readings that were clearly wrong because of the low battery in the sensor. Removing voltage values lower than 2.4 volts looks important as those are the dying batteries, so we go ahead with removing the data rows with the voltage parameter less than 2.4 volts. However, we find that about 30% pf the data has voltage of the battery much higher than 3 volts. They are even higher than 100 volts which is quite strange. There is definitely some other scale or transformation in those voltage values which is unexplainable with our resources and needs further investigation. We move forward assuming that there is error in the information recording and donot discard the 30% of our valuable data.

The work[6]says that the analysis has been done for 44 days in the summer from April 27, 2004 to June 10, 2004. However, we find that there are around 40% of the data entries have date in the month of November which could be seen in the figure 2. We would like to know is the reason for the dates being so anomalous and weather the data recorded on these dates is representative of the summers for which the microclimatic analysis have been performed or not.

After having done with the filtering, we ensure that the data types of all features is the one desired for manipulations and analysis. Moreover, we add certain columns to the dataframe like height/position, direction to help us in the future work. We have also split the result-time which was a combination of the date and the time to just two columns of date and time seperately to ensure easier interpretation and smoother workability with the data.

3.3 Data Exploration

To gauge the data, we first explore the distribution of each of the data in one dimention of value only. The ranges of the data were pretty much, although not entirely in the sensible range of the features they should be proving the correct functionality of the sensor. We can refer to the distribution in figure 3. Our temperature distribution is not as unimodular as the that of the distribution by the authors of the paper[6]

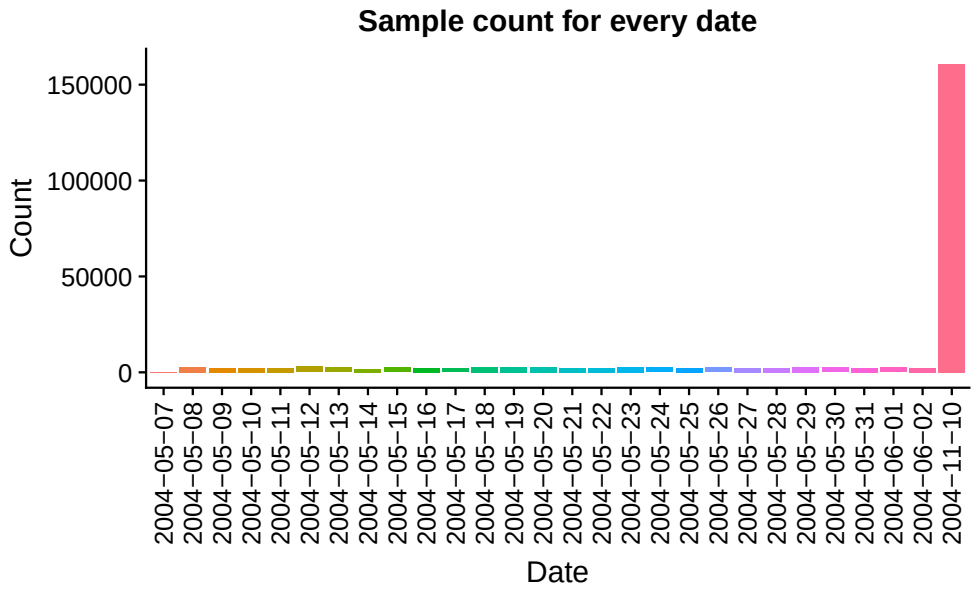

Figure 2: Distribution of the Dates of the Sensor recordings

Figure 3: Distribution of the sensor readings projected onto the value dimension

because of the sudden falling of the values in the range of 15 degree celcius to 20 degree celcius and then a jump after 20 degree celcius. The only data cleaning step that we had performed different than the authors is including the data points with voltage higher than 3 volts. So, in order to ensure that that different step is not the one that leads to this anomaly, I am comparing the density plots from the two data sets in the figure 4. Since both the distribution are not unimodular, the author's analysis is not reproducible with the dataset. The distribution of the humidity shows a bump at 100 RH which is rightly attributed to the fog which is prevalant in the California coast early in the summer. Gladly, the distribution of the incident PAR and the reflected PAR is very similar to that of the author given in the paper's [6] figure3(a).

Exploring furthur, I came across the fact that the time is not representative of the actual time in the sonoma county. This could be seen from the fact that sun rise times during summers of 2004 in the sonoma county according to the verified sources[2] was between 05:45:00 to 06:15:00. However, we can see in the figure 5 that the Incident PAR sensor is activated at the time around 13:00:00 when we see the presence of the presence of preence of the incident PAR values from a long break of when the incident PAR values are 0 which is attributed to no sunlight. Similarly, we observe the sun to betting at about 3:00:00 which contradicts to the records[2] that show it to be setting between the time of 20:00:00 to 20:30:00. Therefore, we can deduce that the timings recorded in the data are off by 7 hours.

During the data exploration, one more thing that we gathered which might be appropriate to note was that majority of readings came only from some nodes/sensors. we find that around 20% of the sensor nodes contribute to around 60% of the data as can be seen in the figure 6. Thus our analysis would be skewed to these nodes position, height, and other conditions. We would not cater to the other microclimatic changes as much as these one. Keeping this inconsistency in mind would help us ask better questions and more insights while doing our analysis. Ignoring this has the potential to cause disastrous impacts on our results down the line.

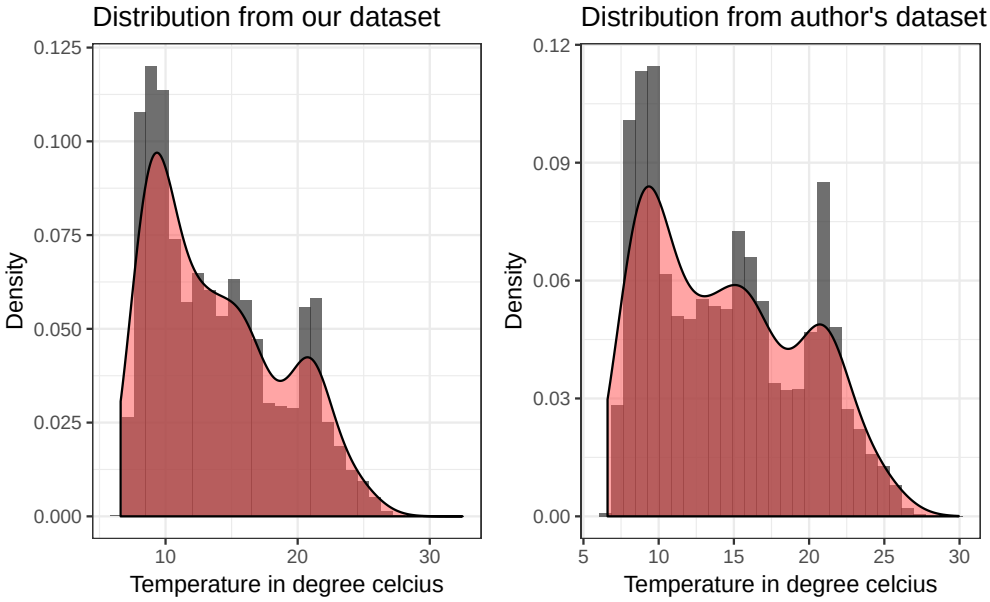

Figure 4: Comparing the temperature distribution from the dataset of our data cleaning vs author's data cleaning

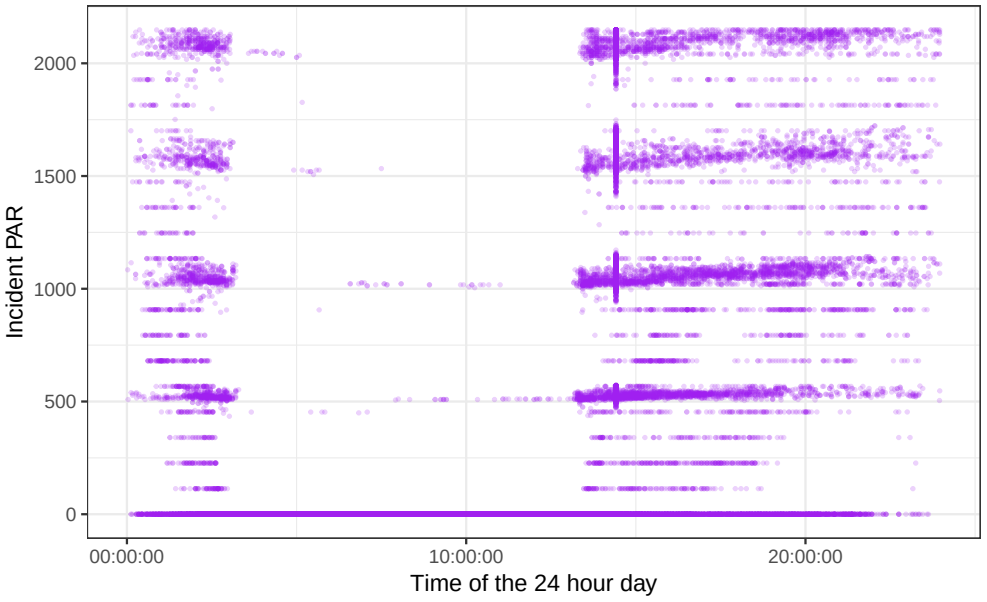

Figure 5: Incident PAR wrt Time

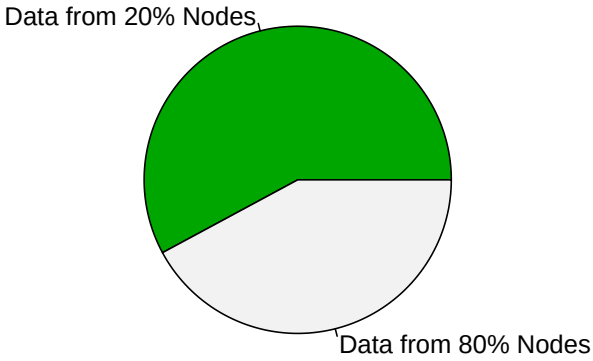

Figure 6: The contribution of few nodes to the majority of data collected

4 Graphical Critique

The authors of the "A macroscope in the Redwoods" paper for their analysis start with "stage 1" of projecting the data to one dimension of value. This one dimension help them to understand the range and distribution of the data independent of the space and time dimension. According to me this should be a data exploration rather than analysis to verify the credibility of the sensors and ensuring that the data is making sense. Also, the work makes statemnet about the humidity being bimodal and the temperature being unimodal. which only is concievable if we look deeply at the histograms and infer them. I feel that a density plot would have made their point more visually clear.

Coming to the the graphs generated of the distribution of the sensor readings projected onto the time and value dimension in the figure 3(b), the distribution of the temperature, humidity, incident PAR and reflected PAR is shown with box plots for each day to see the weather movement in large. The work also successfully explains the correlations seen between the features on a very high humidity day like May 7th. It would have been visually better if the graphs could be made on the common x axis of the day so that we could see the inteplay of the features more clearly. Also, the questuon of the temporal distribution of the climate has not been answered in the text quantitatively. We have not been provided that how much percentage variation we see in temperature, humidity or PAR between say a foggy day and a clear day. In the absence of such numbers the complete picture of the temporal distribution remains incomplete.

Unlike for the temporal distribution graphs of 3(b), I really appreciate that the spatial distribution graphs of 3(c) have been made on a common height axis which makes the interplay between the features more visually comprehensible. The results have been well quantified in the text that what percentage changes in the variables we observe from the top of the tree to the bottom. Also, to account for the lack of spatial trend because of the fact that the amount of variation over time overwhelms the amount of variation over space, they came with a plot of variable's difference from the mean which is figure 3(d). This has been commendable as we can now observe the values realtionship with the height which was hidden in the previous graph 3(c). However, this graph would have been more insightful if the intensity of color would have been used to show the amount and direction of the deflection of the features from the mean with respect to the height.

The combined analysis, figure 4, successfully answers a lot of questions. It explains why the mean incident PAR is higher in the afternoon because of the placement of the sensors in the west. It shows the temperature as the function of the sun and also attributes the solar influence as the reason for the spread in the readings

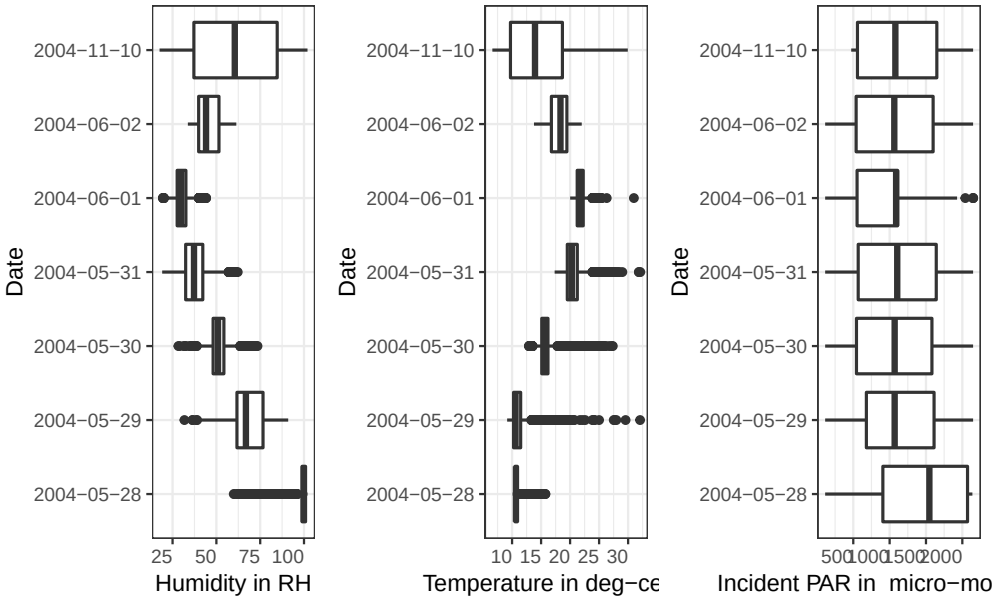

Figure 7: "Distribution of sensor readings wrt time during the later stages of the experiment"

of the sensors. The sudden 50% change in humidity is accounted but not a very convincing explaination for the same is produced. One the sun has risen it is shown that the temporal variability reduces but not the spatial variability. The difference in humidity movements and the temperature movements are justified with the local humidity. The humidity dip observed is termed as the trend exception in the microclimate. These are not convincing enough without being explained by the domain knowledge of the climate.

The spatial analysis (right side of figure 4) is done at the time of the rapid dip observed in humidity that is uncorrelated with the temperature and incident PAR. There is no reason provided for doing so. I would analyse a time of noremal behaviour of variable rather than exception because analysing an exceptional point of time would make the findings of the analysis less credible. The authors must justify the reason of doing so. The presence of correct trend is shown in the graphs. The difference in the temperature between the eastern and western sensors are appropriately captured. The missing part still remains the quantification of the variations in the temperature, humidity and incident PAR with changing time and height.

5 Findings

5.1 First finding

During the temporal trends analysis in the paper [6], the high humidy day May 7th has been discussed. The hypothesis made that it is due to heavy fod has been proved by the low incident PAR skewed lower on the day. The combined analysis pick up the day of May 1st quoting that it has wide range of temperature and humidity readings throughout the day. But there is no analysis of the day on which the humidity is low compared to the nearby days. Also, there is no analysis of the later days of the experiment discussed in the work. That motivated me to look for a day as shown in figure 7 with lower than normal humidity during the later stages of the experimental timeline.

We selected June 01 as the humidity distribution was lowest compared to the other days in consideration. As we expected, the temperature distribution of that day is higher than the rest of the day, proving what authors mentioned in the paper [6] about the inverse relationship between temperature and humidity. However, the incident PAR during that day is not reflecting a positive relationship with the temperature. This can be attributed to the fact that the correlation between the temperature and solar radiations weakens during the

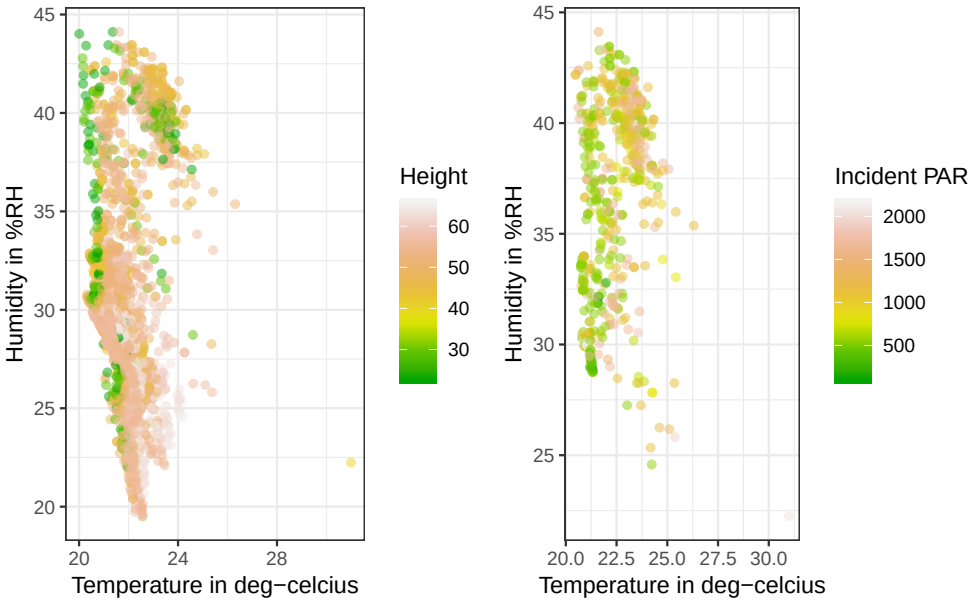

Figure 8: "Variations in Humidity and temperature as a function of (a)Height and (b) Incident PAR"

summer season[1]. Since the summer has now advance in the full form since the time of the initial May 01 analysis, we associate this to the summer anomaly with low correlation of temperature and solar radiation.

We also analyze the variations of humidity and temperature with respect to height on this day of June 01 (Figure 8(a)). We find that at lower heights the humidity is high and the temperature is low. This could be attributed to the canopy at the lower heights in the forest that makes the microclimate damp, moist and cool.

To study the incident PAR, we remove the readings when the radiations are zero i.e. at night because it overwhelms the distribution making it hard to visualize the variations of the radiation during daytime. From figure 8(b) we infer that the radiations readings are weaker when the temperature is lower. Therfore, we deduce that although the relationship of radiations and temperature is not very strong due to summer anomaly on JUNE 01, the temperature is still a function of the radiation and they are positively related however weak.

5.2 Second finding

Sun rises in the east and sets in the west. We wanted to visualize the implications of this universally acceptable truth on the radiations recorded by the sensors of the macroscope in the Redwoods. The paper[6] already mentions how the eastern sensors are warmer in the morning time. However, behavior of the incident PAR with respect to time has not been studied. To analyze the incident PAR with respect to time we chose four days, two chosen near the 1st quartile (May 12-May 13) and two near the 3rd quartile (May 21-May 22) of the time of the experiment. For both the cases we find (figure 9) that the eastern sensors (in red) record the last incident PAR reading before the western sensors. Therefore, it substantiates that when sun sets, the western region recieves radiations for longer. Likewise, we also observe that the eastern sensors start recieving their radiations few minutes earlier to the western ones ratifying that sun rises in the east.

5.3 Third finding

The nodes in the macroscopic network were placed at a variable distance from the trunk of the tree. We divide the sensors to be either far from the trunk if the distance from the trunk is greater than 0.1m or

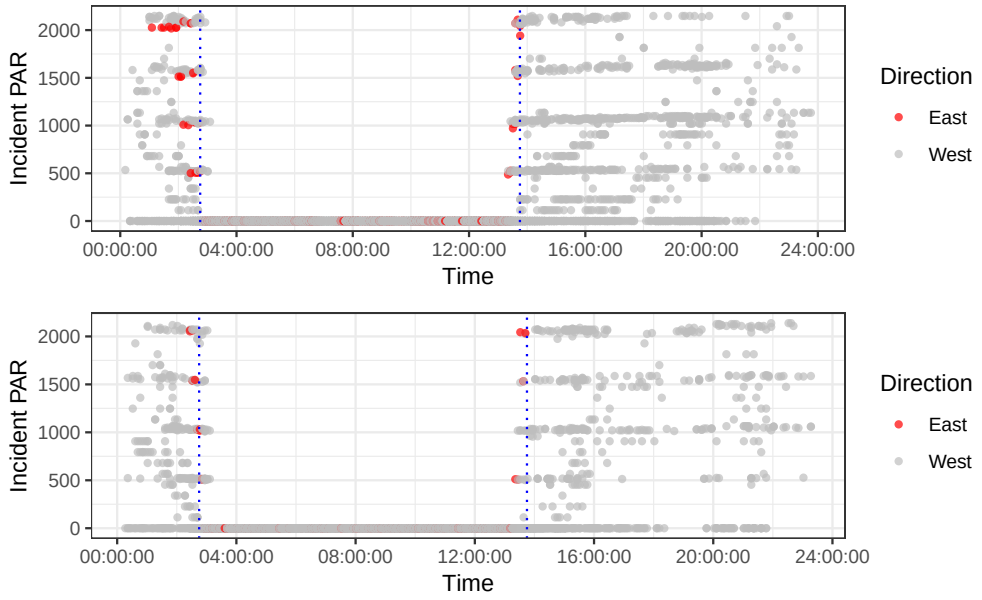

Figure 9: Incident PAR as a function of time (a)Dates : May 12 and 13 and (b) Dates : May 21 and 22

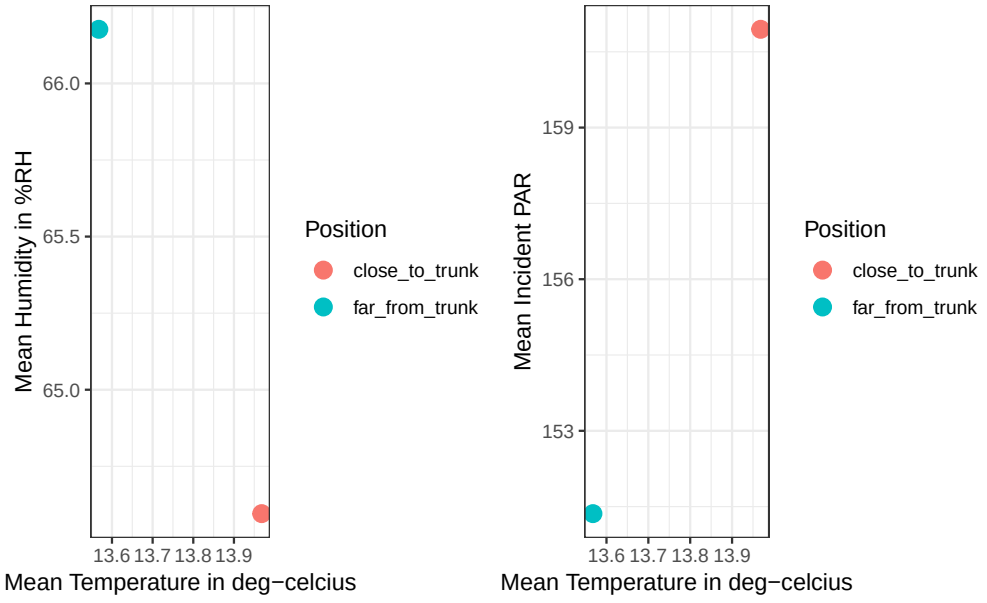

Figure 10: Variables change with respect to distance from the trunk

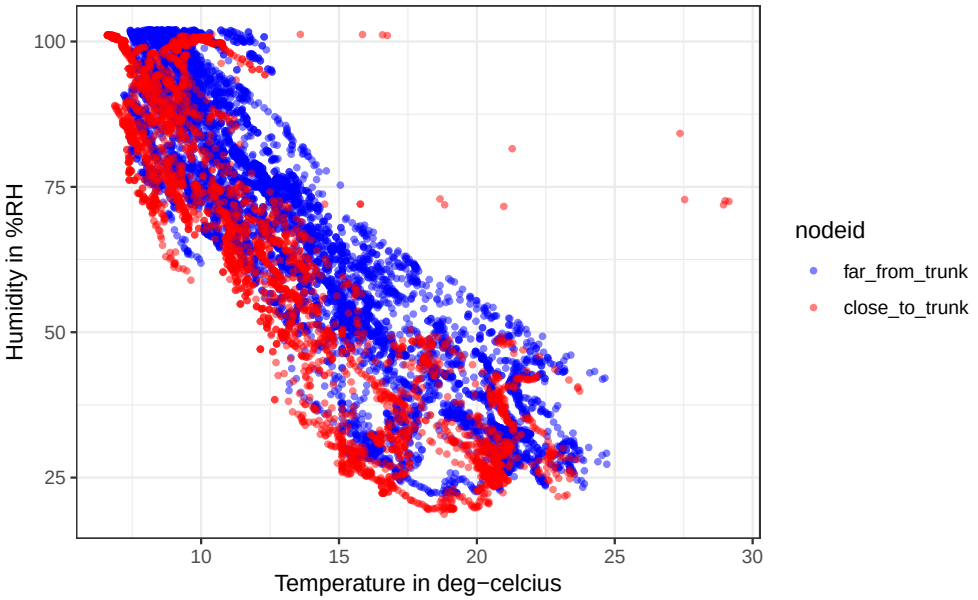

Figure 11: Distribution of temperature and Humidity from a sensor placed close and far to the trunk and each

close to the trunk otherwise. We find that the temperature tend to be 3.7% higher in the sensors far from the trunk because the shade of the canopy is limited and there is direct radiation of sun. The readings of the incident PAR are 6.7%higher in the sensors at edge than those in the interior. Since higher temperature regions can hold more water, for the same amount of water holded the relative humidity goes down. We find the humidity at the edges is about 4% lower than humidity of sensors that are very near to the trunk. This is coherent with our moist and dampy experience in the canopy of the forest.

6 Discussion

Interesting observations are made from the data exploration like the summer anomaly and the relationships between the variables on that anomalous day. We learnt how the solar radiations are different for the different sides of the tree at the same time because of the orbiting direction of earth around the sun. We explored how the overall temperature and humidity varies as a function of the distance from the trunk.

In our analysis, the data size was barrier for a lot of operations. The massivity of the data overpowered the trends and demanded a lot of computational power. Therefore, for most of out analysis, we use a subset of data filtered by either time or sensor or height.

7 Conclusion

The sensor network macroscope offers a useful tool to monitor dense temporal and spatial variations. The work of Tolle et al. captures the complex environmental dynamics of the microclimate surrounding a coastal redwood true satisfactorily. Some of the more peripheral findings from the data that make an interesting observation are discussed in this reports. The challenging work of extracting information from the large amount of data from sensor network deployment is eased with the help of multi-dimentional data analysis.

References

[1] Keith L Bristow and Gaylon S Campbell. On the relationship between incoming solar radiation and daily maximum and minimum temperature. *Agricultural and forest meteorology*, 31(2):159–166, 1984.

[2] abc def. Sonoma, california, usa - sunrise, sunset, and daylength, may 2004, May 2004.

[3] Jason Hill, Robert Szewczyk, Alec Woo, Seth Hollar, David Culler, and Kristofer Pister. System architecture directions for networked sensors. *ACM SIGOPS operating systems review*, 34(5):93–104, 2000.

[4] Samuel Madden, Michael J Franklin, Joseph M Hellerstein, and Wei Hong. Tag: A tiny aggregation service for ad-hoc sensor networks. *ACM SIGOPS Operating Systems Review*, 36(SI):131–146, 2002.

[5] Larry M Miloshevich, Holger Vömel, Ari Paukkunen, Andrew J Heymsfield, and Samuel J Oltmans. Characterization and correction of relative humidity measurements from vaisala rs80-a radiosondes at cold temperatures. *Journal of Atmospheric and Oceanic Technology*, 18(2):135–156, 2001.

[6] Gilman Tolle, Joseph Polastre, Robert Szewczyk, David Culler, Neil Turner, Kevin Tu, Stephen Burgess, Todd Dawson, Phil Buonadonna, David Gay, et al. A macroscope in the redwoods. In *Proceedings of the 3rd international conference on Embedded networked sensor systems*, pages 51–63. ACM, 2005.

[7] Alec Woo, Terence Tong, and David Culler. Taming the underlying challenges of reliable multihop routing in sensor networks. In *Proceedings of the 1st international conference on Embedded networked sensor systems*, pages 14–27. ACM, 2003.