Project1

Xiaoying Liu

12/16/2019

#Abstract The current analysis is based on a case study conducted by Tolle et al. on a wireless sensor network named "Macroscope" in the redwood forest. In the case study, the researchers adopted a creative methodology and deployed the sensor network to monitor the microclimate surrouding a redwood tree, which was not feasible with traditional measurement methodologies.

Base on the raw data collected by the research team, we conduct further statistical analysis through pre-processing, visualization, and data exploration. Our analysis verified some of the findings on the microclimate of redwood tree, in addition, upon analyzing data, we provide possible suggestions of adjustments that can be employed in future data collection methods, as well as possible improvements on the illustrations in the pivotal paper.

#Introduction The research was conducted by scholars from Computer Science Division and Department of Intergrative Biology at University of California, Berkeley, and researchers from Intel Research Berkeley. In the research, the researchers designed a reliable deployment methodology of monitoring dynamic and cross-spatial biological features, and conducted a rigorous experiment with the new method. In the experiment, a redwood tree in redwood forest in Sonoma, California was selected to be the target of measurement. The time period of the experiement is chosen to nearly 44 days during early summer.

#Data ##1.Data Collection Data in the case study is gathered from a 70-meter tall tree for 44 days. A data collection framework including TinyOS and TASK software is implemented. The time period of the experiement is chosen to nearly 44 days during early summer, with a periodiocity of every 5 minutes, to capture most dynamic microclimant variation.

The data collected was identified by two identifier, the label of nodes in the network as *nodeid* and the time the record took place as *epoch*. In the data collected, we are mainly interesed in four variables, tempertaure, humidity, incident PAR (photosynthetically active solar radiation) and reflected PAR. Other variables include voltage readings of the mote sensors, the time of the corresponding record, and the location information of sensors' placement.

#2.Data Cleaning ##Consisitencty check By plotting the histograms of each varibles in two data file, we first find out that the magnitude of variable *Voltage* are completely different for the two datasets. In order to find the mapping relation between voltages from two data sets, we look into 2 data sets and identified same entries using the two identifiers. With the duplicated data, we can now observe how the same data was entered differently. We noticed that they are inversely proportinal to each other.

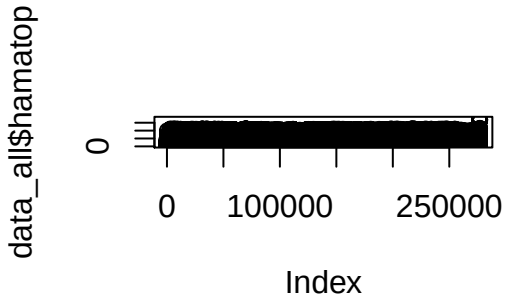
##Missing Values After combining LOG and NET data, we now roughly look at the combined dataset and find out that missing values appear in the four variables of interest. And all missing values are observed to be present spontaneously across the four variables. One possible cause might be the nature of data collection hardware. Once the device failed to record at an epoch, all data will be missing for this entry. Furthermore, most of missing values are located in *epoch* 756 to 1770. As a result, we removed all rows with missing values in the dataset. In total the dataset had 8852 missing values and we now obtained a dataset with 310,179 entries.

##Outlier rejection From the case study we know the motes are numbered from 0 to 200. However, we detected one incorrect entry with *nodeid* 65,535. We then deleted this enty. Since there is only one entry with this error, we don't expect there would be any impact on the data distribution. In addition, according to paper, the researchers stated that the errorous voltage readings were closely linked to wrong readings in other variables. in order to avoid potential influence on our following analysis, we decided to remove voltages which is less than 2.4v or larger than 3v, such measurements totaling 33,123 entries. We now obtained a dataset with 277,056 entries.

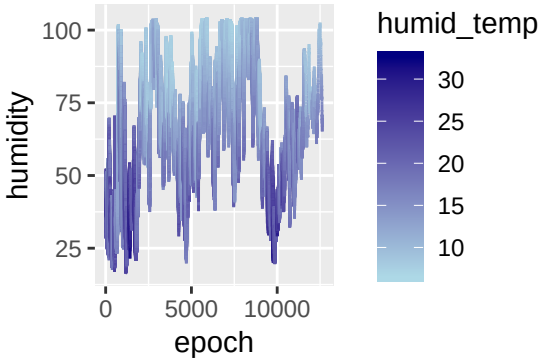
As for humidity, at the begginning of our analysis, we found out that negative humidity readings existed, which violated the physical principles. After above removal of the entries, however, we found these negative reading had already been removed as well. Possible cause could be that negative humidity reading might be relevant to extreme voltage, indicating failing mote. Furthermore, by plotting the scatterplot of all humidity readings across the dataset, we identified that there are 3 other outliers at *nodeid* 118 *epoch* 8717- 8719. The 3 humidity outliers all have value of 114.894, which largely exceed the rest of the dataset whose maximum was merely 104. We then removed the three outliers.

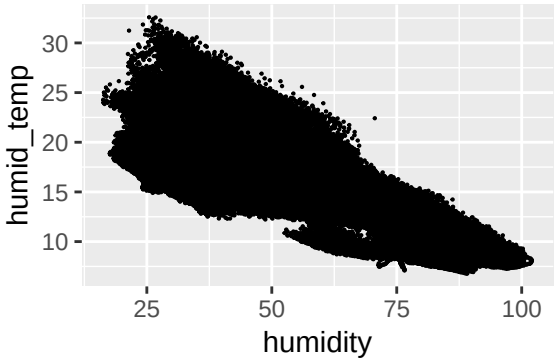
#3.Data Exploration An important discovery of the dataset, when we made a histogram of the observations across *epoch*, was that the number of observations decreased sharply along time at varying rates. We want to look into a reasonable time period to ensure most variation of the nodes in terms of their heights, while the number of observations to be relatively stable.

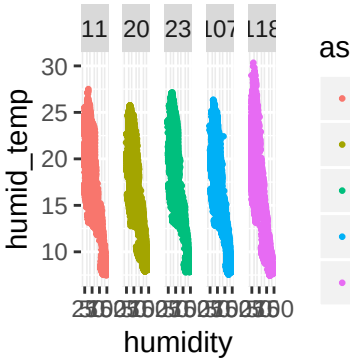
For epoch. we selected epoch 0-2000 and epoch 3000-8000. Firstly, we observed the data points after epoch 2000 decreases sharply and the decrease doesn't slow down until around epoch 2800 to an extend that data points almost stabalize. We therefore determined the first segment to be epoch 0 to 2000, since most nodes of smaller heights still preserves in this segment. Secondly, after epoch 10250, we lost considerable number of datapoints again, with only very limited number of nodes still recording data. We selected epoch 3000 to 8000 as our second segment, since this segment has adequate number of observations, though their distribution was heavily concentrated in higher nodes. We decided to conduct analysis seperately on these two epoch periods, based on the different purposes of our analysis. In addition, we disgarded the last epoch periods. And we need to select certain nodes from the dataset, in 3000-8000, we select nodes which has the most observations for plotting. In the mean time, these observations better have more location variations to avoid confounding factors. Eventually we selected *nodeid* 14, 46, 110, 118, and 119 as our top 5 nodes to do pairwise plots.

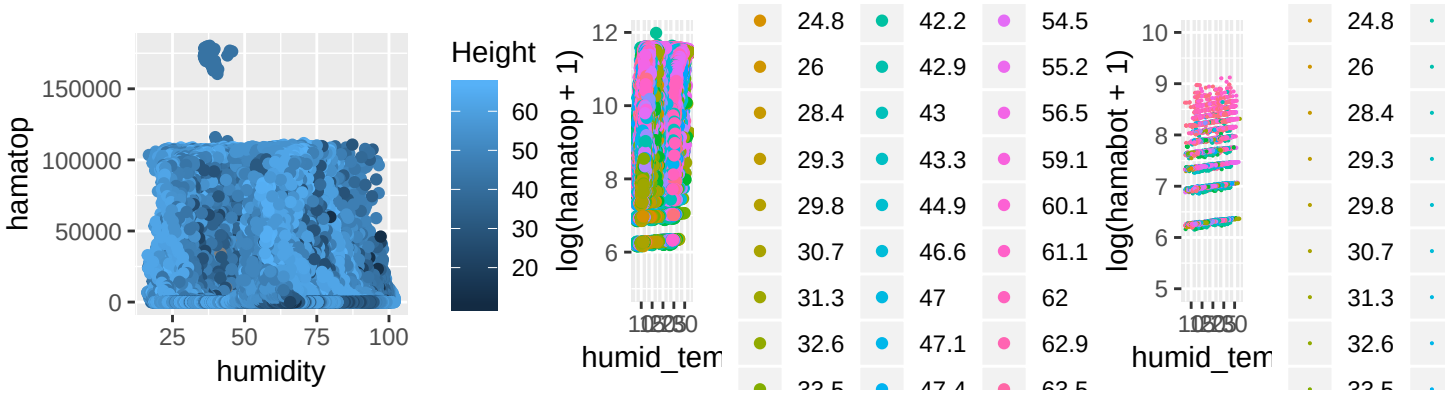


We will do a series of plots: temperature over epoch.(Refer to Fig) . humid-epoch with temp(Refer to Fig), humid-temp(top 5 nodes, 5 plots,refer to Fig). humid-hamatop/hamabot(grid-wise, with height as labels,Refer to Fig)



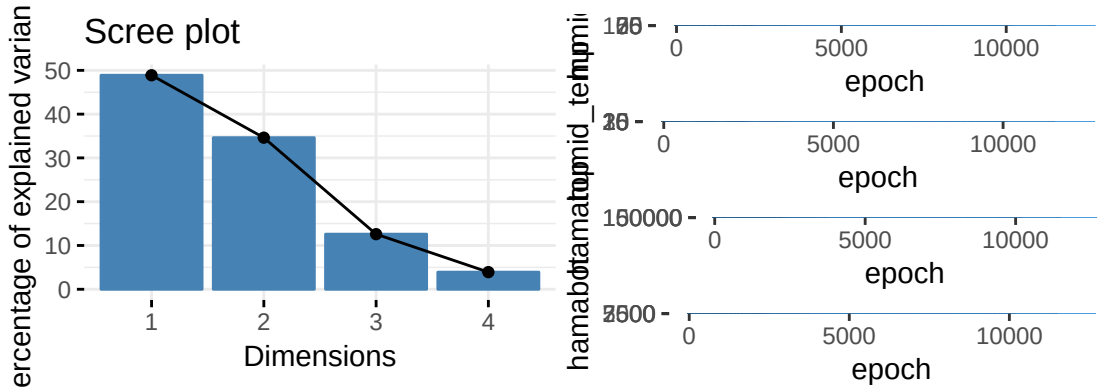




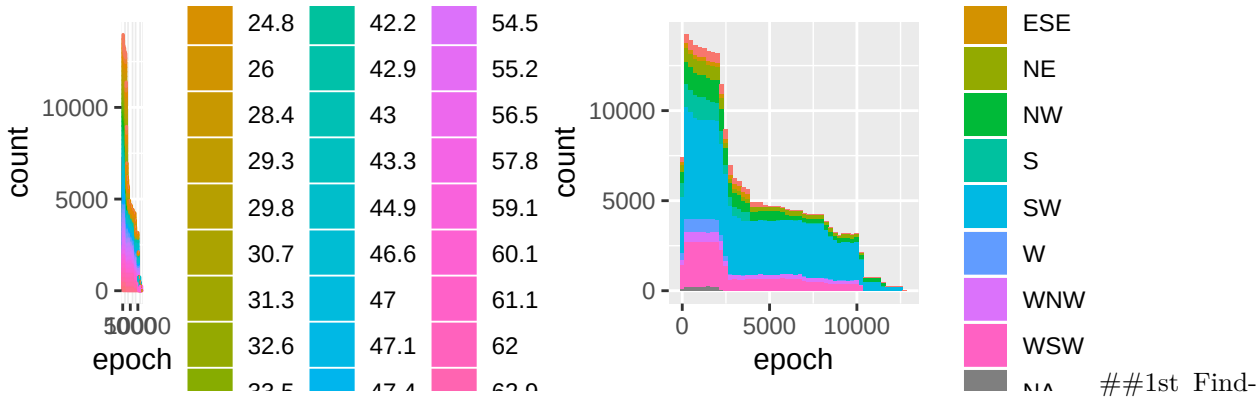


We find out that height, as a predictor is associated with incident PAR. Incident PAR is proportional to corresponding sensor height. Such phenomenon indicates that higher portion of redwood tree would accept higher PAR in general.

we performed PCA(Refer to Fig, scree-plot). From the PCA output, we find out that with the first 2 PCs, the data could be well explained. Thus low dimensional representation is feasible for this dataset.



#Findings



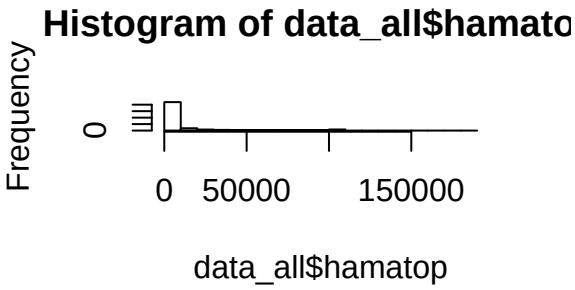
##1st Finding Data collection are highly related to how the sensor system are deployed. In the paper, the researchers addressed their concern about the yield rate of the notes. It is argued to be natural that motes get lost during the process of collection. In our analysis of the amount of readings as epoch increases, we found out that the amount of data collected decreased over time. Furthermore, the pattern of different motes exhibit disparities.

In the figure below we visualized the amount of data across epochs. It is not difficult to see the data drops steadily but at varying rates. We further colored the plot by height, attempting to see whether the height placement of the motes will make them easier to get lost. From the plot we saw a clear pattern that the motes that were lower have a higher rate of not reporting readings. This provides insights for the researchers

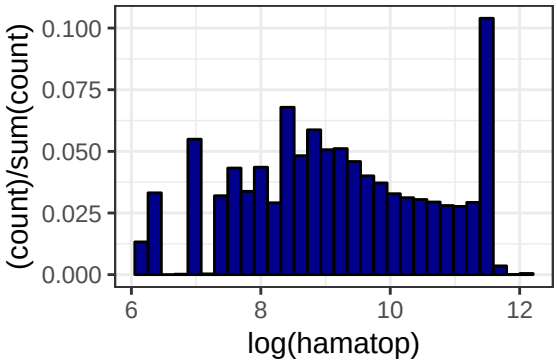
that they may want to improve the set up for motes at lower height or increase the number of these motes in the network.

##2nd Finding We further went to analyze if the data loss is related to the direction of how the motes were placed.Directions include W,SW,WSW,E etc. We produced a similar histogram but colored it by the direction of readings. We witnessed that mote of some directions became completely missing in the latter half of the experiment. This adds to the suggestion that the researchers can strengthen the sensor network by placing more motes to these direction.

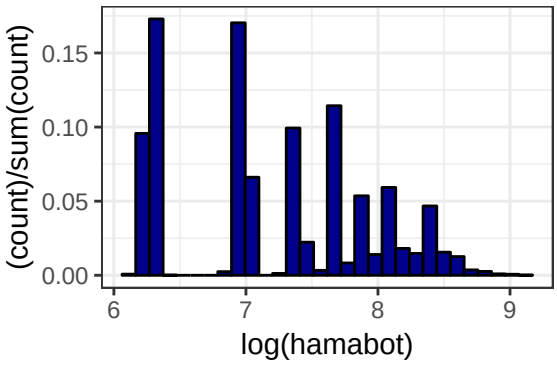
##3rd Finding In analyzing the relationship between temperature and humidity, we observed an overall negative correlation of these two variables. Although it was difficult to conclude if the relation is linear, we can still tell as the temperature appears low, the humidity is more likely to be higher. We also examined different nodes to validates this finding in case that this is only an accidental trend for one node. The analysis confirmed that this negative correlation exists.



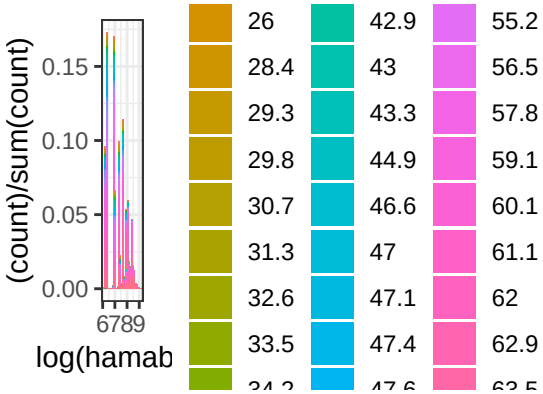

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



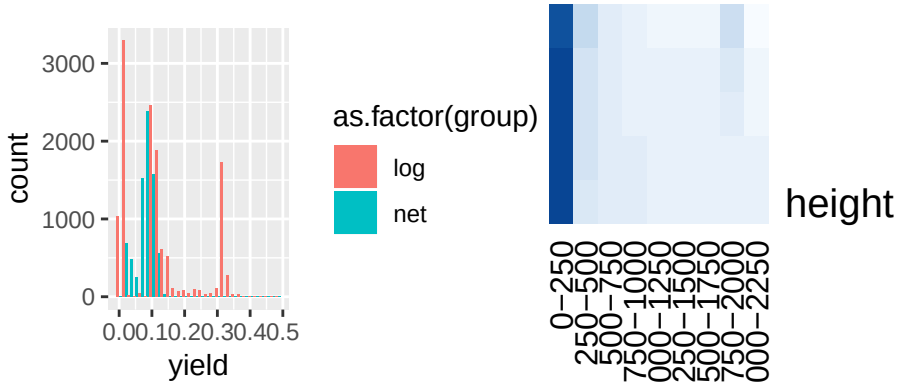

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



#Conclusion From the perspective of data cleaning, data collection is crucial to data quality, we discovered that data quality is affected becasue of battery failure. Lack of data dictionary cost more time to do data cleaning. In addition, since the dataset are large, we believe that hidden informative findings still exist, that if we are equipped with rich domain knowlegde, we could have discovered more.