

# STA 521: Project 1 Redwood Data

Release date: **Thursday, Sep. 08**

Due by: **3 PM, Thursday, Oct. 13**

## Submission instructions

It is a good idea to revisit your notes, slides and reading; and synthesize their main points BEFORE doing the project.

A main report (font size at least 11 pt, less or equal to 12 pages) generated by Latex or Rnw is required. The main report should address the questions below clearly and preferably with figures. The clarity of your writing is also one important grading factor. Arrange the figures compactly if you use .Rnw to generate the report. Keep only the essential plots for the main report. You should aim your writing as smooth as a top research paper. Unlike your other homework, NO CODE should appear in the write-up. A .Rnw file corresponding to the project is also uploaded for you. You may use that to write up your solutions.

You may use the report title “Project 1 Redwood Data Report” or you can also be creative. Put your name (with Student ID) and your teammate’s name (with Student ID) in the **author** line below the title of your report.

The recommended work of this project is at least 16 hours (at least 8 hours / person). Note that it is almost impossible to finish the project in one day. Plan ahead and start early.

You need to submit the following:

1. A pdf of your write-up ( $\leq 12$  pages) to “PROJ1 write-up”. NO Code. Please take care of your writing and figures. 20% of total points will be removed for reports with more than 12 pages.
2. A .tex .R, .Rmd and/or .Rnw file, that has all your code, to “PROJ1 code”.

*Ensure a proper submission to gradescope, otherwise it will not be graded.*

Please read the submission guidelines properly to avoid confusions. Be aware that some of the questions are inherently open ended. Your answers will be graded based on not only the relevance, but also the clarity.

This project allows you to apply previously learned knowledge on data cleaning and data exploration on a real dataset. Here, we focus on data understanding and exploration using appropriate statistical methods and providing well explained visualization of the data, which might be useful for further study. Our work could be considered as an extension of the original paper (Tolle et al.) with statistics and visualization focus.

## 1 Data collection (20 pts)

The data is taken from Tolle et al.. A pdf of the paper can be found on together with problem statement on Sakai Resources. You should read this paper before doing the lab and understand the source of the data. The main data files are packed in **redwooddata.tar.gz**. Take a look at the textfile "read-me" before touching the data. The main data files of interest in this project are **sonoma-data-all.csv** and **mote-location-data.txt**. Explain to your teammate the main conclusion of the paper and how the sensors in the paper work (no need to write for this line).

- (a) Write a summary (1/2 page) about the paper. At least, points such as the purpose of the study, where the data is collected, the main conclusion and impact should be covered.
- (b) Write a summary (1/2 page - 1 page) about the data collection. At least the following points should be covered: How are the sensors deployed? What is the duration of the data recording? What are the main variables of interest? What is the difference between the data in **sonoma-data-log.csv** and that in **sonoma-data-net.csv**.

## 2 Data cleaning (40 pts)

This data set is quite raw - it contains some gross outliers, inconsistencies, and lots of missing values. Read the **Outlier rejection** section in the paper carefully and critically. You don't have to blindly follow their data cleaning method.

The file **sonoma-data-all.csv** is a simple concatenation of the two files **sonoma-data-log.csv** and **sonoma-data-net.csv**. However, doing the merge of two data files requires that they are consistent. nodeid and epoch

together provides a unique identifier for one measure. But some other variables are not consistent.

- (a) Check histograms of each variable in two data files (Plot only the ones that you think are interesting or relevant). Which variable is not consistent? Convert the data to the same range. NO CODE but explain clearly what you did.
- (b) Remove missing data. Comment on the number of missing measurements and the corresponding date and time period.
- (c) The location data is separate in another file **mote-location-data.txt**. Incorporate it in the main table. Hint: here the nodeid serves a key to add columns for height, direction, distance and tree. State the number of variables in your new data frame.
- (d) Use histogram and quantiles to visually identify easy outliers for each of the four variables: humidity, humid temp, hamatop, hamabot. And remove them. Comment on the rationality behind your removal.
- (e) (Bonus) Discuss other possible outliers and explain your reason why it is better to remove them than to keep them.

### 3 Data Exploration (40 pts)

- (a) Make some pairwise scatterplots of some variables. Pick a reasonable time period. Explain your choice and describe your findings.
- (b) Are any of the predictors associated with Incident PAR? If so, explain the relationship.
- (c) Each variable of our data basically have three dimensions: value, height and time. Consider each variable as a time series and look at its temporal trend. Generate such plots (value vs time) with height as color cue for at least four variables (Temperature, Relative Humidity, Incident PAR and Reflected PAR). You can do it for different time scales (during an hour, during a day or during the entire experiment). However, at least the plots with days as x-axis are required. Comment on the range, continuity and strange behaviors in these variables.
- (d) After PCA analysis, generate scree plot of the data. Can this data be approximated by some low-dimensional representation?

## 4 Interesting Findings (15 \* 2 pts)

Describe two/three interesting findings from exploratory analysis of the data. Try to use the techniques that you have learned, such as histograms, PCA, K-means, GMM and hierarchical clustering etc. Note that even though you got a dataframe with only a few columns, you may reshape the dataframe before doing any EDA, such as reorganizing such that aggregated information in each day is a column, or a particular hour in each day is a column. Comment on your interesting findings. Different bonuses are given based on how interesting your result is.

- (a) Finding 1.
- (b) Finding 2.
- (c) (Bonus) Finding 3. Bonus is given only if we find all three findings interesting.

## 5 Graph Critique in the paper (40 pts)

The overall quality of the paper by Tolle et al. is good. However, some plots are not perfect from a statistician's point of view.

- (a) Figure 3[a] shows the distributions of sensor readings projected onto the value dimension, using a histogram. It turns out that both the incident and reflected PAR have long tail. We could not read full information from this histogram. Try to make a better plot with log transform of the data.
- (b) What message do the boxplots in Figure 3[c] and 3[d] try to convey? Do you think the plots convey the right messages? If not generate a new plot with the same data. Hint: compare to some plots in Figure 4.
- (c) Any suggestions for improving the first two plots in Figure 4? Can you distinguish all the colors in these two plots?
- (d) Comment on Figure 7. Is it possible to generate a better visualization to highlight the difference between network and log data?