Stat 222: Sensor Project

1 Data Description

For this project, your primary data source will be from a wireless sensor network, which captured spatial and temporal information (e.g., temperature, humidity) from the microclimate around a coastal redwood tree [1, 2]. Wireless sensor networks (WSNs) are becoming increasingly prevalent in a growing number of areas including health care, environmental and industrial monitoring, as well as home automation.

WSNs are typically composed of hundreds to thousands of nodes connected by a wireless network topology (e.g., star or mesh). Each node is composed of a radio transceiver and antenna, a microcontroller (running a minimal, embedded Operating System such as TinyOS), one or more sensors, and a power source (e.g., a battery or a way to harvest solar, kinetic, or thermal energy). A central challenge (or goal) of WSN design is to create low cost, tiny sensor nodes. This constrains resources such as the battery, memory, computational capacity, and communication bandwidth. Often energy is the scarcest resource. WSNs are designed to be resilient to node failure, capable of withstanding harsh environmental conditions, and easy to deploy and use.

A central goal for this project is to give you hands-on experience working with the types of messy, incomplete, and inconsistent data that you will encounter in real world applications. It is also designed to give you more time to develop your ability to understand and critique statistical graphics as well as to provide you an opportunity to practice designing good graphics to convey information and reveal patterns. Finally, the project will provide you with the opportunity to practice reading and writing about applied statistical data analysis.

2 Your Assignment

This is an individual project. While you may talk to other students in the class about the assignment, you will be responsible for producing your own work. This includes all code, figures, and text.

You are free to use either R or Python.

Exploration of Data

Your first task will be to check the data quality. This involves understanding the data collection method as well as all the potential data entry issues (e.g., missing values, errors in data). Please read the paper to understand how the sensor works, and write a paragraph to discuss the measurement of each variable you find interesting in the data. Please have at least 3 variables (related to your findings) in your report.

Bearing the data quality in mind, your second task will be data cleaning. 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 will need to do some cleaning of the data but do **not** blindly follow their method. Record in your report the steps you take and any evidence (i.e., summary statistics and EDA plots) you use to support them.

Next, think of some questions you would like to ask of the data and use R or Python to answer them graphically. Try to show what interesting findings can be gained from the data. You may show general patterns or anecdotal events. Using the entire dataset may be challenging. Try just a subset of sensor nodes or a day's worth of data. You may also need to jitter (i.e., adding a small amount of randomness) to your data, so that you don't overplot or overlap elements. Again record in your report your process—include plots you make. Don't be afraid to try methods that are new to you and be critical of your own graphics.

Graphical Critique

Critique the plots in Figures 3 and 4 of the original paper. You should make sure to incorporate the material that Deb Nolan presented in class as well as the assigned readings. In particular, you should carefully consider and address the following questions:

1. What is the data? What observations and what variables are included (or not)?
2. What is the message? What questions does the graphic try to answer? Does the graphic answer them successfully? Does it raise any questions not addressed in the text?
3. How would you improve it? Be specific. Discuss both minor tweaks that would improve on the existing graphic as well as alternative graphics including possibly additional data, which may be better suited for the questions at hand.

Presenting findings

Choose three interesting findings from your exploratory data analysis and produce a publication quality graphic for each along with a short caption of what each shows. I expect to see very polished graphics. Think carefully about your use of color, labeling, shading,

transparency, etc. Again, you should be sure to review the material from the readings and the guest lecture.

Timeline and logistics

Here is the tentative schedule:

Monday	Wednesday
(2/8) Start Sensor Project	(2/10) Social networks
No class	(2/17) TBD
(2/22) Twitter posters	(2/24) Financial data
(2/29) Sensor report	

We will **not** discuss this project in class. This project will require a significant effort on your part. You will need to understand the paper as well as the data. Identifying and dealing with data quality issues will involve careful investigation and thought. Since you will be simultaneously working on your Twitter project, you will need to be careful with your time and will need to communicate clearly with your Twitter project team about deadlines and work schedules.

3 Report Details

I've created a Git repository for you with a directory with the following structure:

```
sensor
|-- data
|   |-- mote-location-data.txt
|   |-- README.md
|   |-- sonoma-data-all.csv
|   |-- sonoma-data-log.csv
|   |-- sonoma-data-net.csv
|   `-- variable_key.txt
`-- report
    |-- Makefile
    |-- sensor.bib
    `-- sensor.tex

2 directories, 9 files
```

The main files of interest are `sonoma-data-all.csv` and `mote-location-data.txt`.

I have also provided a template in your individual class repositories for the write up. Since the template is intended, in part, to make grading easier, please do not deviate from it without good reason. Please restrict your write up to twelve pages, including figures. This is a strict limit.

I expect your reports to be professional. Among other things, this means you should carefully prepare and revise your report as well as the graphics you generate for it. The quality of your writing and graphics will be included in your grade for this assignment.

References

[1] Gilman Tolle, Joseph Polastre, Robert Szewczyk, David Culler, Neil Turner, Kevin Tu, Stephen Burgess, Todd Dawson, Phil Buonadonna, David Gay, and Wei Hong. A Macroscope in the Redwoods. In *Proceedings of the 3rd International Conference on Embedded Networked Sensor Systems*, pages 51–63. ACM, 2005. http://delivery.acm.org/10.1145/1100000/1098925/p51-tolle.pdf.

[2] Sarah Yang. Redwoods go high tech: Researchers use wireless sensors to study california's state tree. *University of California at Berkeley News (UCNEWS)*, 2003. http://www.berkeley.edu/news/media/releases/2003/07/28_redwood.shtml.