Stat 133, Spring '08
HW #2
Due Friday February 15

Location Estimation of a Wireless Client

For this homework, your goal is to re-format a data frame and perform exploratory data analysis on the transformed data. The data are part of a study to develop a statistical model for tracking people and things inside a building using WiFi technology set up for Internet access. That is, it is thought that the physical location of a client can be estimated from the received signal strength of the client's laptop from five stationary access points. These data consist of received signal strengths from five access points for a set of clients where the true location of the client is also known.

The main goal of this assignment is for you to gain facility in working with data structures in R and to express some creativity in your exploration of the data.

Background: Inside a building, Global Positioning Systems do not provide an effective means for locating people and other things. Instead, systems are built to use WiFi set up for Internet access (IEEE 802.11b) and multiple access points.

In a lab setting, the received signal strength follows a log-normal distribution, and signal strength decays linearly with the log of distance from the access point. But, physical characteristics such as walls, elevators, and people add noise to the received signal strength. Statistics can be helpful in developing a model to predict location from the received signal strengths at other locations.

The data were collected are from an infrastructure-based deployment, where a sniffers (access points) monitor signal strength from clients and uses these to figure out where the clients are located. An alternative is to use client-based deployment, where the client locates itself based on the signal strength it receives from the known access points. The client-based approach has disadvantages compared to the infrastructure-based deployment in relation to power and software requirements on the clients, but it has advantages in terms of privacy.

Data: You are provided with measurements of received signal strength taken at five access points for 254 locations. You are also provided the (x, y) coordinates of these 254 locations. The data are available on the web from
`http://www.stat.berkeley.edu/users/nolan/stat133/data/wireless.txt`
This file includes the locations and received signal strengths of the five access points. These are the final five lines in the file.

Read Data Read the data into an R data frame, and determine its dimension, row names, and column names, if any.

Access Point Locations Remove the information for the five access points from the data frame, and store the x, y coordinates for the five access points in a separate data frame called AP.

Reformat the Data Frame Reformat the remaining information in the data frame such that the five columns for signal strength are collapsed into one column, called "ss". This new data frame should still have the variable "x" and "y" for the locations of the client. In addition, create three new variables. One of these should be a factor indicating which access point measured the signal strength; a second variable should be a factor that indicates which client emitted the signal; and the third variable is the distance from the client to the access point. Note that you may use simple integers to denote access points and clients, i.e. access point 1 through 5, and client 1 through 254.

Exploratory Phase: Before embarking on developing your estimator, explore the data using graphics and statistical summaries. Consider the following questions:

- Is the distribution of signal strength roughly the same for the five access points? And, is it log-normal, i.e. when you look at the log of signal strength, does it's distribution appear approximately normal?

- Do these signal strengths follow the properties described above, as being linear in log-distance?
- Are there any anomalies in the data?

To answer these questions, consider making density plots or normal quantile pltos of signal strength or a transformation of signal strength. Also consider making scatter plots of the signal strength against distance and log distance. Try fitting a line to the data using the R function `lm`. Note that if you save the output from `lm` in an R object, that the object is a list with elements that include the residuals and the fitted values. If the relationship is roughly linear then a scatterplot of the residuals against distance or the fitted values should show no discernable pattern.

Some R commands that might help you are, plot, density, lm, abline, boxplot, hist, points.

Submit For this assignment you are to submit commented code in a **plain text** file as an attachment to your bspace submission. The name of the file should be your SCFUserName.R, e.g. s133cu.R Included in the file should be comments (these begin with #). The comments should indicate your name and SCF account, they should document your code, and they should briefly describe your findings from the exploratory phase of the assignment.

The file that you submit should be runnable in R. That is, from within R you should be able to "source" in the code and comments and not get any errors. In other words, your file should look something like the file, trafficRestructure.R, posted on bspace.