WiFi Fingerprinting

Michael Simko

06/03/2018

Introduction

Determining outdoor location using global positioning satellite signals is virtually ubiquitous. In most instances, cell phones are capable of locating positions within less than 5 meters (~16 feet) https://www.gps.gov/systems/gps/performance/accuracy/#how-accurate. GPS, however, is a line of sight technology and does not work well, if at all, inside buildings. WiFi Fingerprinting is a technology to determine the position of a mobile device based on information collected from a wireless local area network (WLAN) system.

The goal of this analysis is to develop models capable of using signals from wireless access points (WAPs) to locate a position of a user's device indoors. This position will be defined as a combination of building, floor, space ID and relative position.

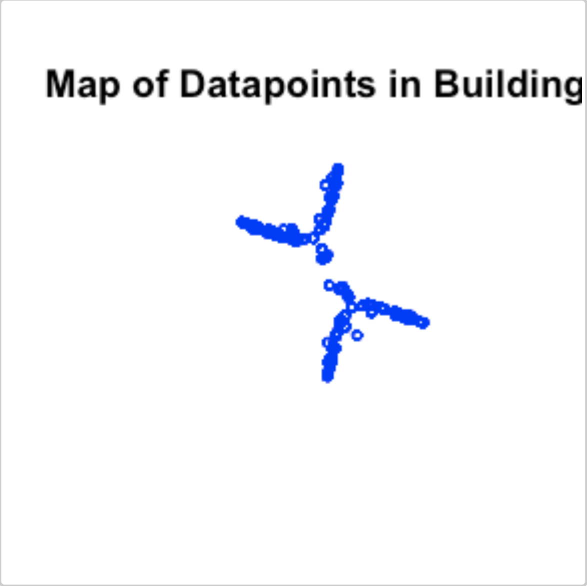
The dataset consists of signal strength (in dB) from 520 WAPs paired with information such as User ID, Phone ID, known longitude and latitude at each sample location, specific reference information (like building, floor, relative position) and a timestamp. Values of +100 indicate no signal, and signal strengths range from 0 (very strong) to -100 (very weak). In this dataset, values of +100 mean no signal. The full dataset is available at http://archive.ics.uci.edu/ml/datasets/UJIIndoorLoc.

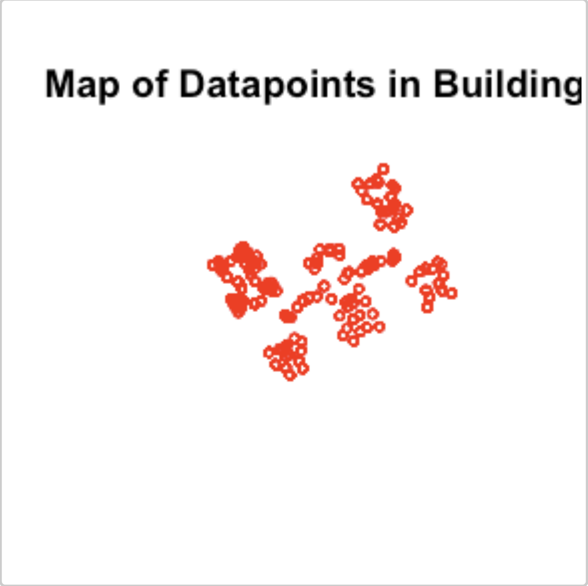
For this analysis, the data will be split by building and further analyzed according to WAPs with the strongest average dB levels (meaning that WAPs with very weak signals at given unique locations will be eliminated from the analysis). The literature states that building design, WAP location and other factors (signal reflection, attenuation) can influence signal strength at given locations, thus affecting the validity of factors that are being used for the model.

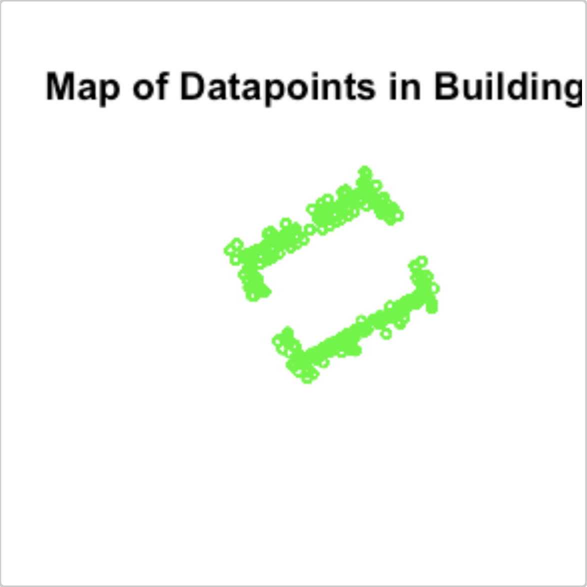
The structure of models made here will be based on signal strength from individual WAPs and a unique categorical factor that maps to each individual study location point.

Longitude and Latitude values were given for each measurement point. These values were mapped to show the location points which sketched out the rough shapes of each of the buildings.

Plot Building Maps





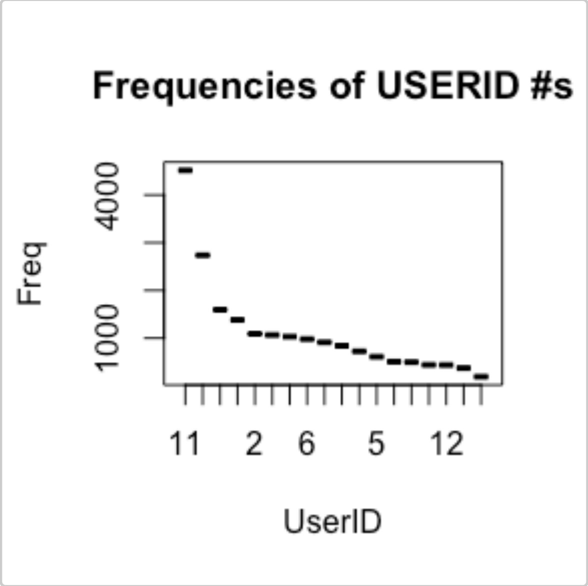


Analyze USERID information

```
## [1] "Table I - Frequency of USERID by Building # and Floor #"
```

	User ID#																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Building																	
0	2737										2512						
1		472		374			740	204	506	362	606		401	447		658	19
2		619	192		610	980	643	303	560	551	1398	437	440	1149	498	374	52

	User ID#																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
#Total cases	2737	1091	192	374	610	980	1383	507	1066	913	4516	437	841	1596	498	1032	72



Inspection of the USERID datapoint frequency table by building shows a few interesting patterns. First, most users reported data for more than one building, however, User #1, for instance, only collected datapoints in Building 0 (and a large number of datapoints compared to the rest of the set too). Also noteworthy, User #11 was the only user to collect datapoints across all three buildings, and also collected the largest number of total datapoints. User #3 collected the fewest number of total datapoints and only in Building 2.

Focus on USERID - Building and Floor

```
## [1] "Table II - Frequency of User#11 by Building # and Floor #"
```

	Floor			
	0	1	2	3
Building				
0	531	599	710	672
1	139	467		
2			785	613
#Total cases	670	1066	1495	1285

```
## [1] "Table III - Frequency of User#1 by Building # and Floor #"
```

	Floor			
	0	1	2	3
Building				
0	528	757	733	719
#Total cases	528	757	733	719

```
## [1] "Table IV - Frequency of User#14 by Building # and Floor #"
```

	Floor			
	0	1	2	3
Building				
1	138	309		
2			610	539
#Total cases	138	309	610	539

```
## [1] "Table V - Frequency of User#7 by Building # and Floor #"
```

	Floor	
	0	1
Building		
1	550	190
2		643
#Total cases	550	833

```
## [1] "Table VI - Frequency of User#2 by Building # and Floor #"
```

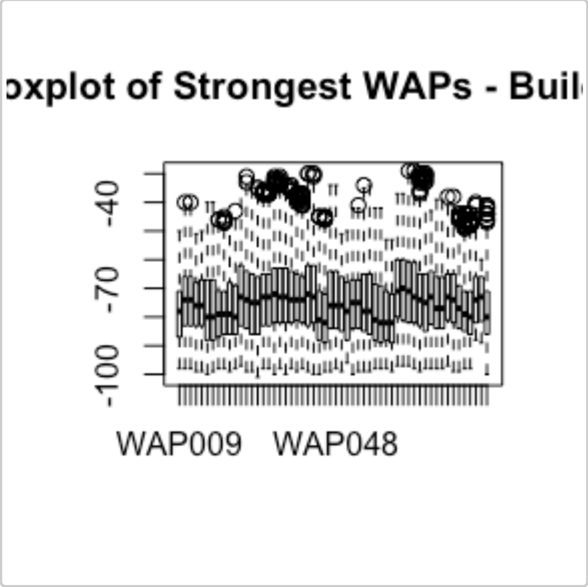
	Floor	
	2	3
Building		
1	472	
2		619
#Total cases	472	619

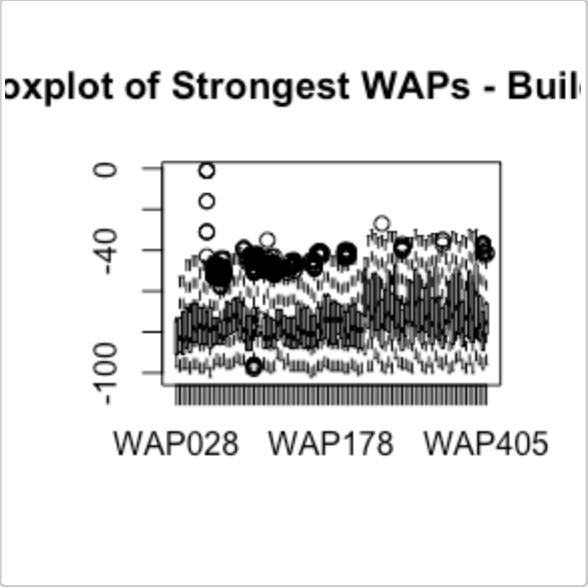
Feature Engineering

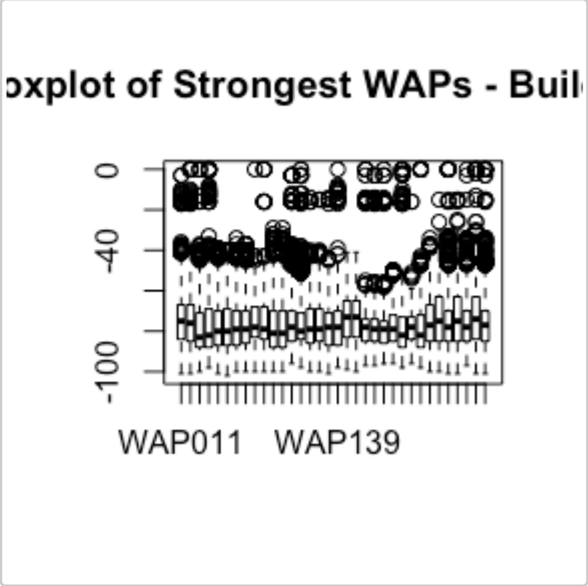
To refine the analysis, new dataframes are created with only the strongest WAP signals. This includes removing all location information other than the unique position vectors and only keeping those WAPs with average signal strength > -80 dB.

Graphical Analysis of WAP signals

Boxplots are used to illustrate the number of WAPs measured in each building with average signals greater than -80dB.







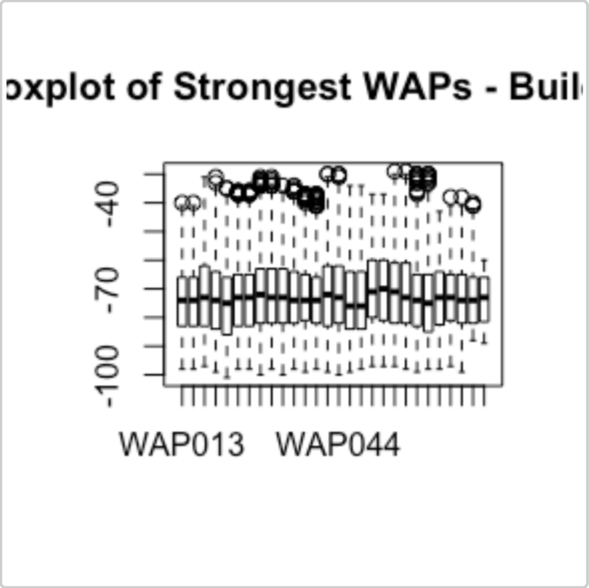
```
## There are  56  WAPs with average signals > -80 dB in Building 0.
```

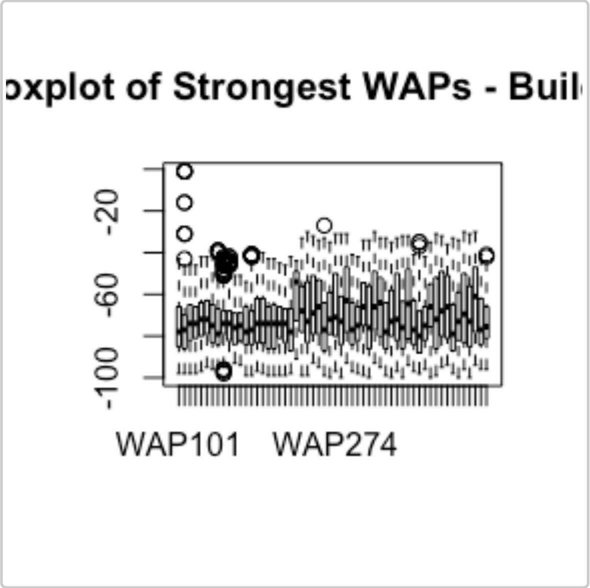
```
## There are  93  WAPs with average signals > -80 dB in Building 1.
## There are  34  WAPs with average signals > -80 dB in Building 2.
```

Inspection of these boxplots shows that in Building 0, there are no WAPs with individual signal strength levels greater than about -30. In one instance for Building 1 and in multiple instances in Building 2, groupings of individual WAP signals exceed -20 and some reach as high as 0 in some cases. This seems to suggest that some signal levels in Building 2 are unusually high, or most of the signal levels in the other buildings, for some reason, are lower than can be expected.

Further reduce datasets

In order to further simplify the modelling process, new threshold values for minimum WAP signals in each building are selected. The goal is select a signal threshold to reduce the number of strongest points to between 20 and 50. Building 2 needs no further processing, only Buildings 0 and 1 need to undergo more pruning.





```
## There are now  28  WAPs with average signals > -80 dB in Building 0.
```

```
## There are now  56  WAPs with average signals > -80 dB in Building 1.
```

Prepare and Run Models for Building 0

Three models were created and run for each of the buildings. The models found to be the most useful are k Nearest Neighbor, CART and random forest. The results for each of these models are shown below.

k Nearest Neighbor for Building 0

```
##  Accuracy     Kappa
## 0.6177175 0.6161560
```

CART for Building 0

```
##  Accuracy     Kappa
## 0.6687949 0.6674437
```

Random Forest for Building 0

```
##  Accuracy     Kappa
## 0.6983240 0.6970929
```

Prepare and Run Models for Building 1

k Nearest Neighbor for Building 1

```
##  Accuracy     Kappa
## 0.5121753 0.5089185
```

CART for Building 1

```
##  Accuracy     Kappa
## 0.5308442 0.5276315
```

Random Forest for Building 1

```
##  Accuracy     Kappa
## 0.5633117 0.5602627
```

Prepare and Run Models for Building 2

k Nearest Neighbor for Building 2

```
##  Accuracy     Kappa
## 0.5912698 0.5899986
```

CART for Building 2

```
##  Accuracy     Kappa
## 0.6732804 0.6722499
```

Random Forest for Building 2

```
##  Accuracy     Kappa
## 0.7288360 0.7279894
```

For each of the three buildings, the Random Forest models generated the best accuracy and kappa values. CART did slightly better than knn in every case, but neither were as good as Random Forest. Building 1 had the lowest modeled agreement values, while Building 2 showed the strongest agreement.