Indoor Positioning Data Exploration

STAT4/510 Wasabees Group

2023-11-30

1. Describe your data
2. Describe the basic variable component of your data
3. Report the various findings you have established so far, with interpretation (include discussion on what you find and add how useful it is to your project objective)
4. Discuss any challenges you encountered, and ways by which you handled these.

1. Introduction - Data Description

Indoor position systems (IPS) development is an active area of research that can be used in numerous settings. Part of efforts to develop, calibration and model these systems are achieved through the use of WIFI signals. The following report, describes and characterizes a large data set compiled in a 15 by 36 meter area that contains six (wifi routers) access points, signal strength, various locations and orientation of the devices (10 parameters total) The data is subdivided in two sub-sets, one denominated "offline data", which corresponds to various testing devices connected to the network at different locations and orientations, and the other an "online data", where 60 locations and orientations of the devices were selected at random.

The offline data was collected designing a 1 meter resolution grid, resulting in 166 locations. In each of these locations, the device was oriented starting at 0 degrees inclination and at 45 degrees increments, and the strength signal measured. Furthermore, each combination of location/orientation was sampled 110 times. This grid sampling is intended to be used to calibrate a indoor positioning model. On the other hand, the online data was designed to simulate real-world data, in which locations are not bounded by the 1 meter grid used in the offline data, and were selected at random. This randomization included the orientation of the device and therefore, the online data consists of 60 randomly selected location/orientation combinations sampled 110 times.

More details of the floor plan, and location of online and offline data can be seen in Figure 1.

For simplicity, this report will share the results found in the offline dataset, but initial process of data cleaning can be directly applied to the online data as well. The online and offline data sets share the same structure so that's why in this document, we explore the offline data set with the expectation to apply the same method for the online data set.

Circles serve as markers for the positions where "offline" measurements were conducted, while black squares indicate the locations of six access points. These reference positions provide a calibration of signal strengths within the building, forming the basis for constructing a model to predict the whereabouts of a hand-held device when its location is unknown. The hand-held device supplies x and y coordinates, akin to latitude and longitude on a map, along with its orientation. Signal strengths are recorded at eight orientations in 45-degree intervals. For every location and orientation combination, 110 signal strength measurements were documented for each of the six access points.

2. Data Processing

In this section, we provide a brief description of steps undertaken to format and clean the data set.

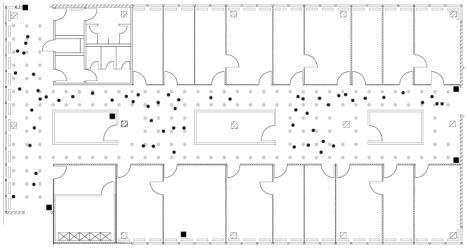


Figure 1: Flooplan location. Access points are squares. Grey dots are offline data locations and black dots are online data locations.

2.1 Variable Description

According to documents provided by the client, the data contains the following variables:

- `time`: time in miliseconds.
- `scanMac`: IP address of the scanning device.
- `pos`: the 3-D coordination of the scanning device.
- `orientation`: the scanning device's orientation.
- `mac`: the IP address of the access points.
- `signal`: signal strength in dBm.
- `channel`: the channel frequency.
- `type`: type of device (access point = 3, device in adhoc mode =1)

2.2 Data Formatting

The data is stored in a .txt file. The first six rows of the data is printed below

```
## [1] "# timestamp=2006-02-11 08:31:58"
## [2] "# usec=250"
## [3] "# minReadings=110"
## [4] "t=1139643118358;id=00:02:2D:21:0F:33;pos=0.0,0.0,0.0;degree=0.0;00:14:bf:b1:97:8a=-38,2437000000
## [5] "t=1139643118744;id=00:02:2D:21:0F:33;pos=0.0,0.0,0.0;degree=0.0;00:14:bf:b1:97:8a=-38,2437000000
## [6] "t=1139643119002;id=00:02:2D:21:0F:33;pos=0.0,0.0,0.0;degree=0.0;00:14:bf:b1:97:8a=-38,2437000000
```

Below are our observations that shaped our approach for data formatting:

- The first three rows, marked by the character #, provides information for the next batch of 110 readings. Similar rows like this re-appear multiple times in the data set.
- The data rows contains a series of variables and values, separated by semicolons. We performed a simple split using semicolon as the separator, and found that some variables, such as `pos` and `mac` are further subdivided. Each `pos` value corresponds to a set of x, y, and z coordination. Each `mac` value corresponds to readings of `signal`, `channel`, and `type` respectively.

Since our model is supposed to rely on wifi signal strength to predict device location, we need to format our data so that each observation corresponds to one single `signal` value.

```
[1] "# timestamp=2006-02-11 08:31:58" "# usec=250"
[3] "# minReadings=110"
```

```
[1] "t=1139643120075"                     "id=00:02:2D:21:0F:33"
[3] "pos=0.0,0.0,0.0"                     "degree=0.0"
[5] "00:14:bf:b1:97:8a=-38,2437000000,3" "00:0f:a3:39:e1:c0=-54,2462000000,3"
```

To transfer the data into a dataframe, we performed the following operations:

- We start by cleaning the rows with the "#" symbol. A total of 5312 rows are removed, resulting in 146,080 rows.
- Second, we utilize semicolon, colon, and equal sign as separators for the different variables in each row and re-format the data so each row represents an observation of the variable signal strength.
- Lastly, we bind all rows together to create a dataframe and enter the proper names for each variable.

We provide the structure of our data frame, along with the first 3 observations below.

```
##            time           scanMac posX posY posZ orientation               mac
## 1 1139643118358 00:02:2D:21:0F:33  0.0  0.0  0.0         0.0 00:14:bf:b1:97:8a
## 2 1139643118358 00:02:2D:21:0F:33  0.0  0.0  0.0         0.0 00:14:bf:b1:97:90
## 3 1139643118358 00:02:2D:21:0F:33  0.0  0.0  0.0         0.0 00:0f:a3:39:e1:c0
##   signal    channel type
## 1    -38 2437000000    3
## 2    -56 2427000000    3
## 3    -53 2462000000    3
```

2.3 Data Cleaning

Before further exploring and analyzing the data, we do a pre-exploration process, to assess if variables should be removed from the data set, and if conversions were required. The list below presents the operations we performed:

- We observe that some variables, such as `time`, `position`, `orientation`, `signal` and `channel`, should be converted to numerical for analysis.
- We now focus on the `time` variable. According documents provided by the client, time is expressed in milliseconds from midnight on January 1st, 1970. We convert this time value to seconds and then designate the class of the time element to visualize the values as date-times in R. Additionally, we retain the more precise time information in 'rawTime' in case it becomes necessary for future analysis.
- Based on the documents received from the client, a value of 1 for the variable `type` corresponds to ad-hoc devices. However, for the development and testing of the IPS, we will utilize only the signals measured at fixed access points. We removed all rows that have `type` = 1. After this removal, the variable `type` has a value of 3 for all observations, so we eliminate the variable `type` from the data set.

Table 1 provides our cleaned and formatted offline dataset.

Table 2 provides a basic exploration of each variable and calculate the mean for the numerical variables (i.e., position, orientation, signal). We find that position-z, has a mean of zero. Further exploration shows that the variable has a value of zero for all the observations in the offline dataset. This seemingly anomalous value is due to the fact that all of the readings were taken on one floor of the building. We were tempted

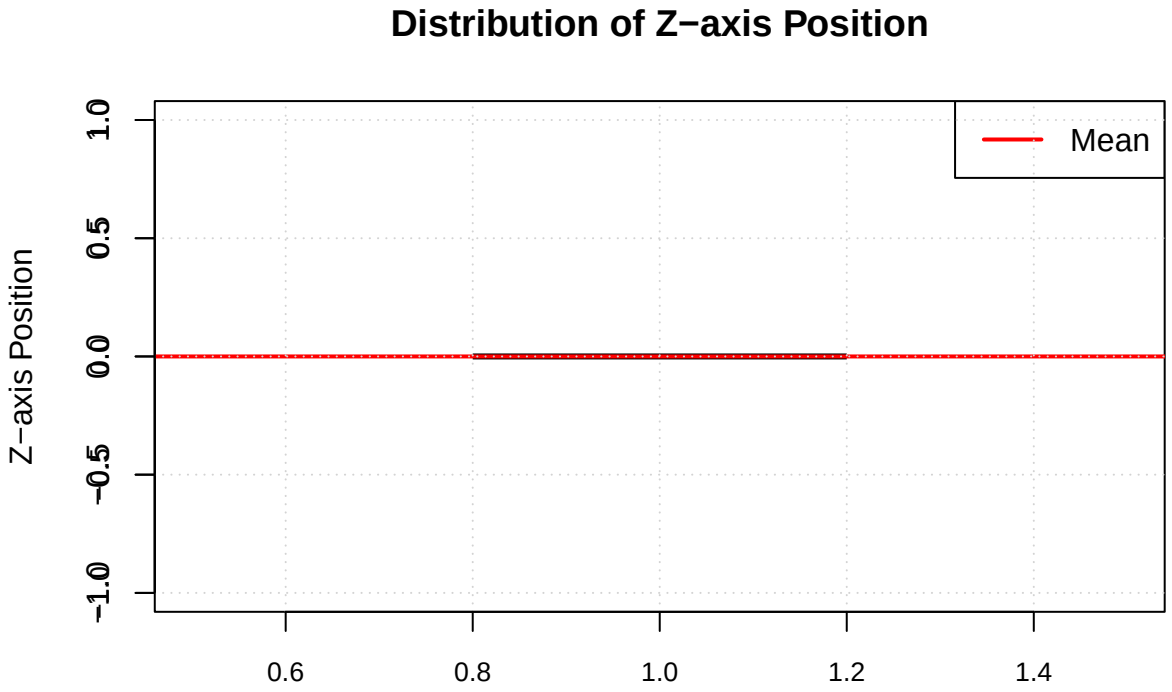
Table 1: Clean Data with transformed variables

time	scanMac	posX	posY	posZ	orientation	mac	signal	channel	rawtime
2006-02-11 02:31:58	00:02:2D:21:0F:33	0	0	0	0	00:14:bf:b1:97:8a	-38	2.437e+09	1.139643e+12
2006-02-11 02:31:58	00:02:2D:21:0F:33	0	0	0	0	00:14:bf:b1:97:90	-56	2.427e+09	1.139643e+12
2006-02-11 02:31:58	00:02:2D:21:0F:33	0	0	0	0	00:0f:a3:39:e1:c0	-53	2.462e+09	1.139643e+12
2006-02-11 02:31:58	00:02:2D:21:0F:33	0	0	0	0	00:14:bf:b1:97:8d	-65	2.442e+09	1.139643e+12
2006-02-11 02:31:58	00:02:2D:21:0F:33	0	0	0	0	00:14:bf:b1:97:81	-65	2.422e+09	1.139643e+12
2006-02-11 02:31:58	00:02:2D:21:0F:33	0	0	0	0	00:14:bf:3b:c7:c6	-66	2.432e+09	1.139643e+12

Table 2: Mean values for numerical variables

	colMeans(mean_off)
posX	13.517162
posY	5.896623
posZ	0.000000
orientation	167.162523
signal	-61.703083

to delete the variable posZ; however, since it is a meaningful variable and we do not know, for now, if the online dataset has different posZ values, we will keep it for the current stage of our project.



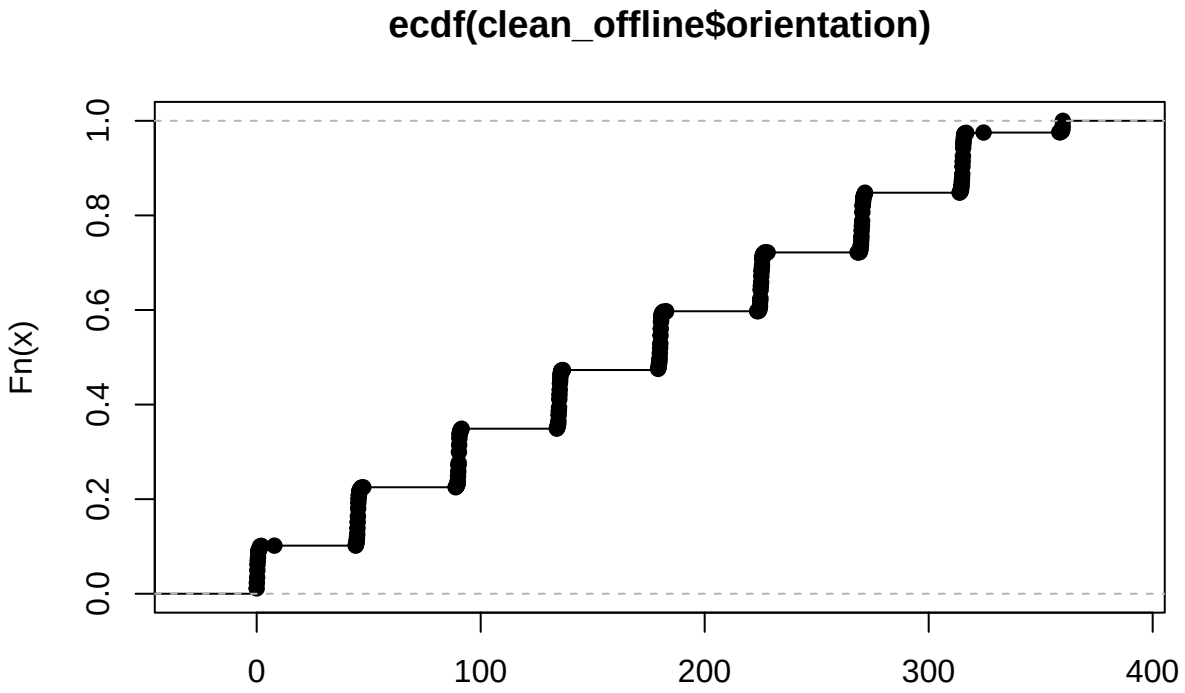
3. Data exploration

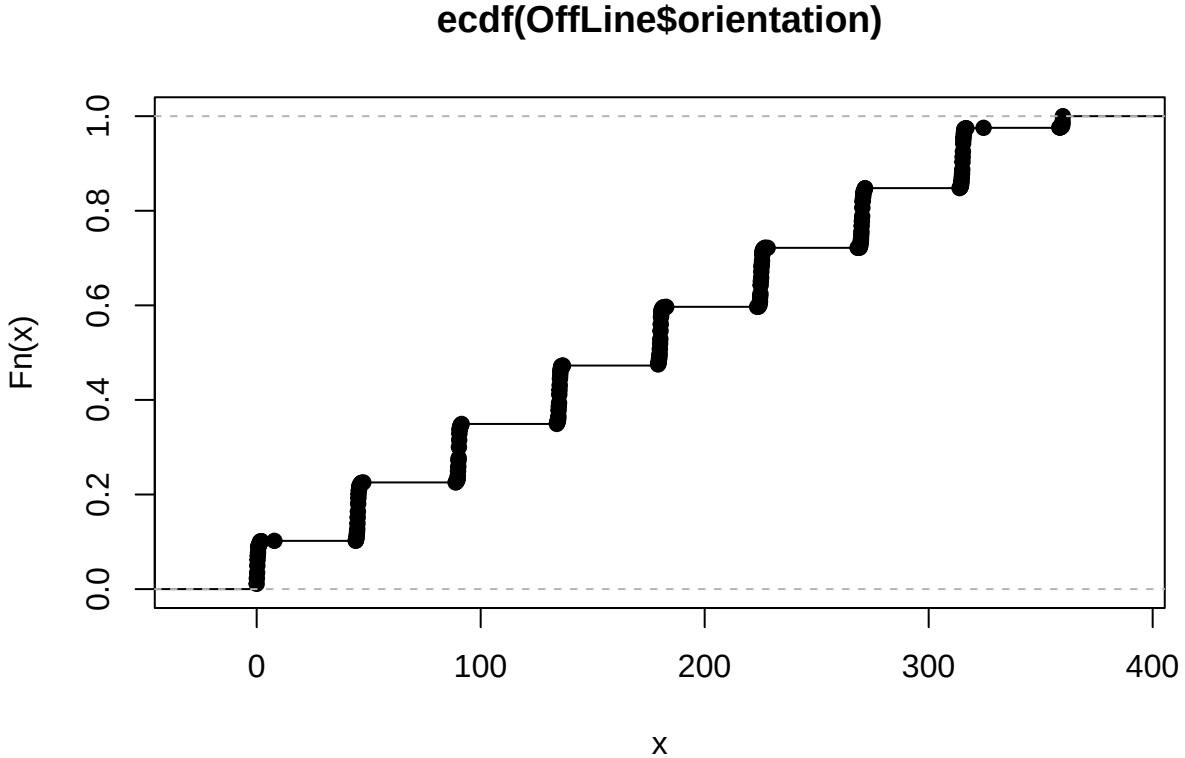
If you remember, the observations for orientation were configured at intervals of 0 degrees, 45 degrees, 90 degrees, and so forth, resulting in more than eight values. To further examine the distribution of the

Table 3: Min, Max and Mean values for posZ

meanposZ	maxposZ	minposZ
0	0	0

'orientation' variable, we will analyze it through an empirical cumulative distribution function (ECDF).





From the plot, we observe a concentration of observations around 0, 45, 90 degrees, and so forth. However, there is evident dispersion in between, with instances like 47.5 degrees, 358.2 degrees, and so on. This diversity in values is not a drawback; in fact, it could be valuable in its current form. Alternatively, we could derive value by categorizing these values into bins to align with the original eight values. To implement this, we will develop a function.

4. Challenges