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.

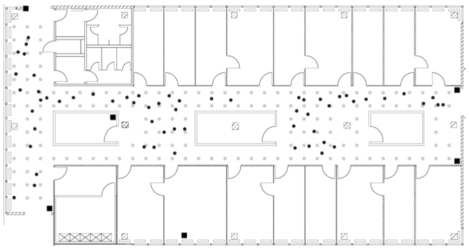


Figure 1: Flooplan location. Access points are squares. Grey dots are offline data locations and black dots are online data locations.

2. Variables Description

2.1 Data Cleaning

The data provided by the client, 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)

As a first goal, prior to exploration, we organize and clean the dataset. In the analysis, various data setup approaches were considered, and a table detailing the variables in the dataset slated for analysis is provided below, with a brief overview to avoid delving into excessive detail.

The format of the data is stored in a .txt file. Each location, and its subsequent 100 samples, are marked by the presence of a "#" symbol with three outputs. Below, an example of a row found in the dataset after some basic cleaning. We observe some variables, such as position ("pos") and mac are further subdivided.

```
[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"
```

Upon further exploration, we create a function to process the data into a suitable dataframe. We start by cleaning the rows with the "#" symbol. A total of 5313 rows are removed, resulting in 146,079 rows. Second, we utilize the semicolon as a separator for the different variables in each row, and account for the variables that were further subdivided (position and mac). Lastly, we add the variable names to the clean dataframe.

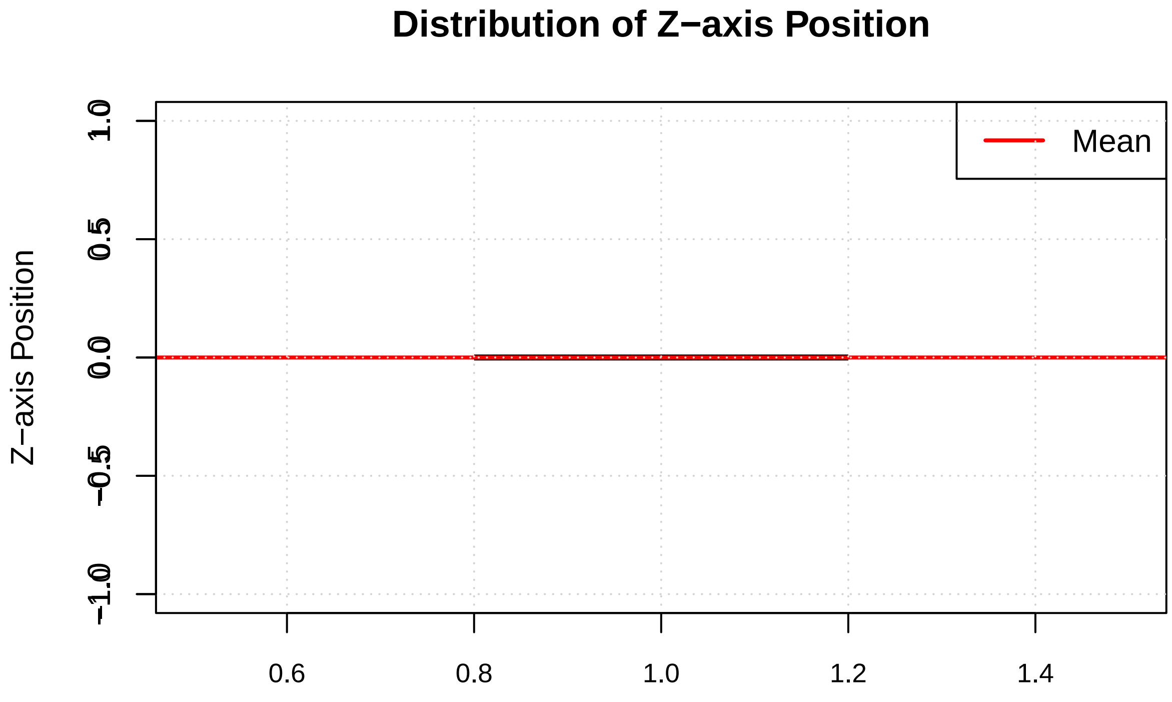
```
##            time           scanMac posX posY posZ orientation               mac
## 1 1139643118358 00:02:2D:21:0F:33  0.0  0.0  0.0         0.0 00:14:bf:b1:97:8a
## 2 1139643118358 00:02:2D:21:0F:33  0.0  0.0  0.0         0.0 00:14:bf:b1:97:90
## 3 1139643118358 00:02:2D:21:0F:33  0.0  0.0  0.0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

Table 1: Clean Data with transformed variables

time	scanMac	posX	posY	posZ	orientation	mac	signal	channel	rawtime
2006-02-10 23:31:58	00:02:2D:21:0F:33	0	0	0	0	00:14:bf:b1:97:8a	-38	2437000000	1.139643e+12
2006-02-10 23:31:58	00:02:2D:21:0F:33	0	0	0	0	00:14:bf:b1:97:90	-56	2427000000	1.139643e+12
2006-02-10 23:31:58	00:02:2D:21:0F:33	0	0	0	0	00:0f:a3:39:e1:c0	-53	2462000000	1.139643e+12
2006-02-10 23:31:58	00:02:2D:21:0F:33	0	0	0	0	00:14:bf:b1:97:8d	-65	2442000000	1.139643e+12
2006-02-10 23:31:58	00:02:2D:21:0F:33	0	0	0	0	00:14:bf:b1:97:81	-65	2422000000	1.139643e+12
2006-02-10 23:31:58	00:02:2D:21:0F:33	0	0	0	0	00:14:bf:3b:c7:c6	-66	2432000000	1.139643e+12

2.2 Formatting

Before further exploring and analysyng the data, we do a pre-exploration process, to assess if variables should be removed from the data set, and if conversions were required. We observe that some features would be beneficial to have as numerical variables (i.e., position, orientation and signal). Furthermore, we check the documents recieved by the client and determine that a value of 3 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 and determine to eliminate the rows that have a type 3 value, and subsequently, eliminate the variable from the data set. We now focus on the 'time' variable. According to the documentation, 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.



We conduct a basic exploration of each variable and calculate the mean for the numerical variables (i.e., position, orientation, signal) and find that position-z, has a mean of zero. Further exploration shows that the variable has a value of zero for all the data, so decide to remove it. This, seemingly anomalous value, is due to the fact that all of the readings were taken on one floor of the building.

Table 2: Mean values for numerical variables

	colMeans(mean_off)
posX	13.732986
posY	5.876238
posZ	0.000000
orientation	167.159037
signal	-63.853949

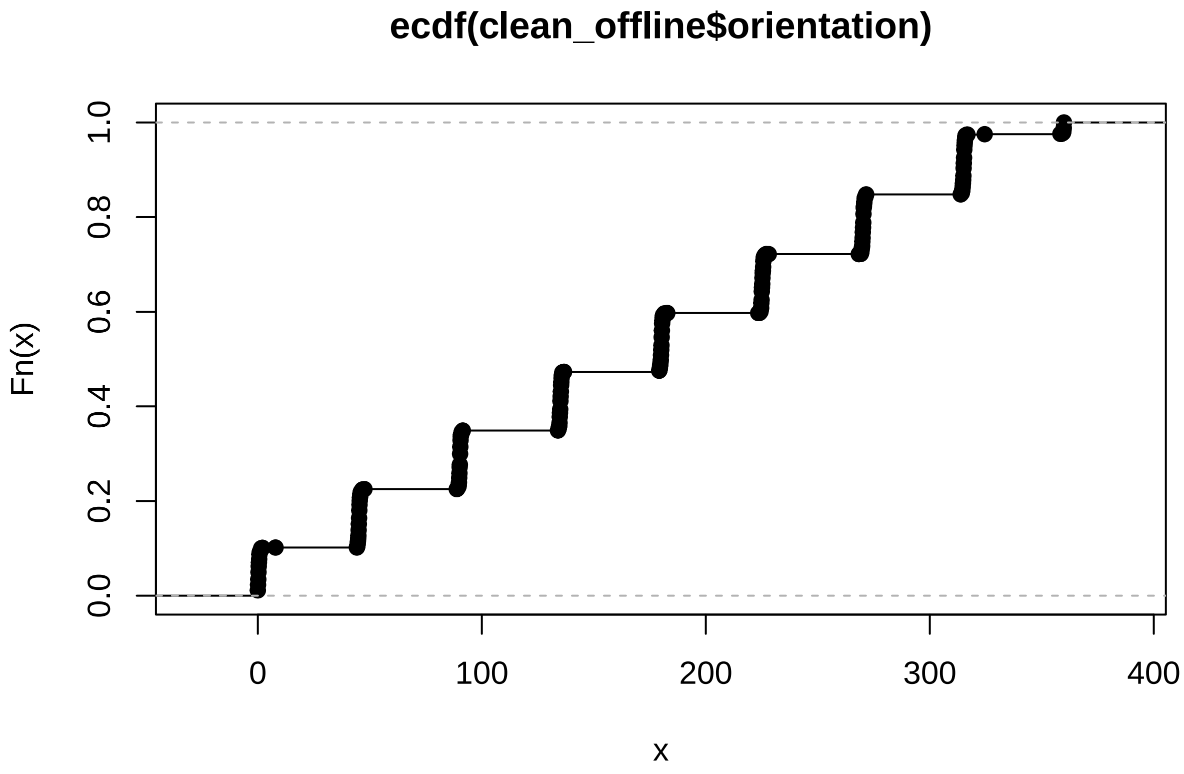
Table 3: Min, Max and Mean values for posZ

meanposZ	maxposZ	minposZ
0	0	0

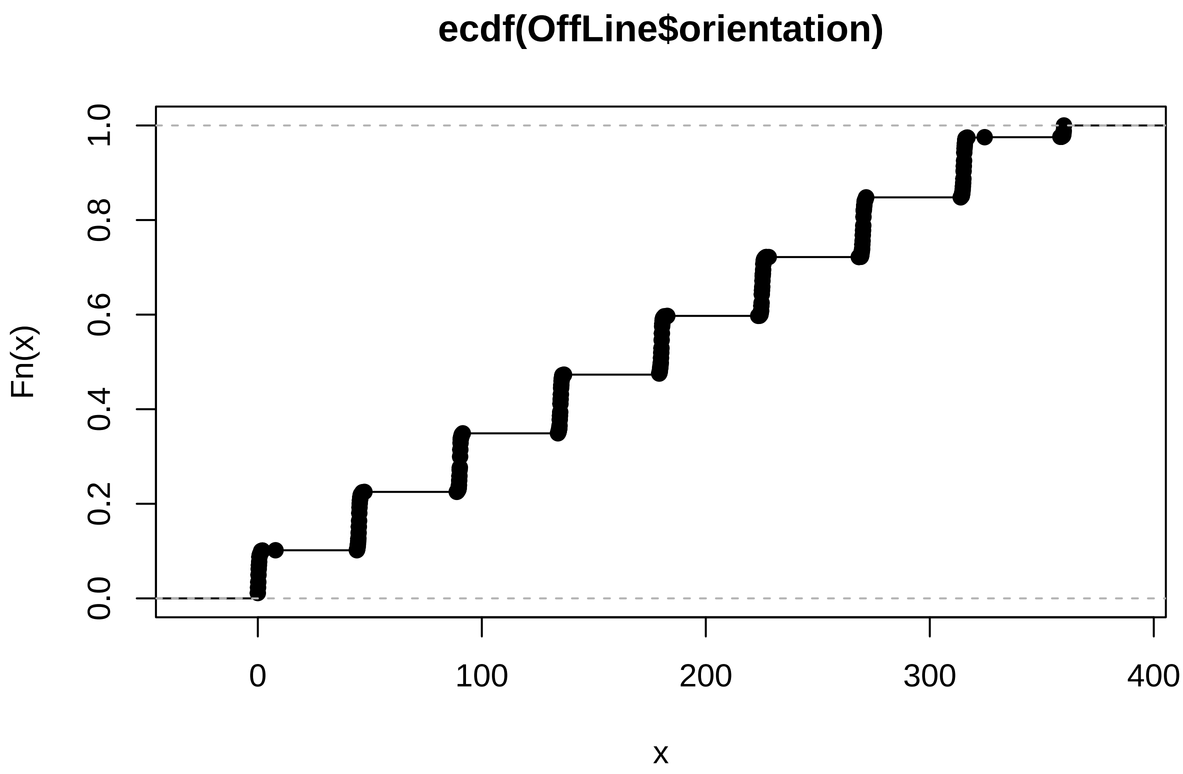
3. Data exploration

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

```r
# Plot the orientation of the measurement device
plot(ecdf(clean_offline$orientation))
```



```
plot(ecdf(OffLine$orientation))
```



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