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 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.

2.1 Variable Description

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

- `time`: time in miliseconds since midnight 01/01/1970 UTC
- `scanMac`: IP address of the scanning device.
- `pos`: the 3-D coordination of the scanning device (x,y,z)
- `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 are stored in a .txt file separated by both hash symbols and semicolons. A sample of the initial format of the data can be seen 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
```

We note a pattern in the internal organization of the data that we use to properly format it.

- The first three rows, are characterized by the # symbol and mark the limit between each combination of location/orientation (i.e., each location sampling is delimited by 3 hash symbols).
- 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.

```
[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"
```

Since the model is supposed to rely on wifi signal strength to predict device location, we organized the data so each observation corresponds to one `signal` value by performing the following operations:

- First, we removed the rows with the "#" symbol that mark the beginning of a location sampling/orientation sample. As a result, a total of 5312 rows were removed, resulting in 146,080 rows.

- Second, we utilized semicolon, colon, and equal sign, as separators for the different variables in each row and re-formatted the data so each row represents an observation of the variable signal strength.
- Lastly, we binded all rows together to create a data frame. Subsequently, we entered the proper names for each variable.

A sample of the first 3 rows is depicted below.

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
```

Before further exploring and analyzing the data, we conducted a simple assessment to convert variables into the correct types defined in thedocumentation provided by the client. We performed the following operations:

- We converted the variables `time`, `position`, `orientation`, `signal` and `channel`, to numerical values.
- We then transformed 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. We retained 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`. This operation removed 203,185 observations, resulting in a new data set of 978,443 observations for all access point devices. We then eliminate the `type` variable (now equal to 3 in the entire data set).

A sample of the formatted data set can be seen in Table 1.

3 Data exploration

This section focus on the exploration of the data itself.

Table 2 provides a basic exploration of each variable and calculate the mean for the numerical variables (i.e., position, orientation, signal).

Table 1: 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	2.437e+09	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	2.427e+09	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	2.462e+09	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	2.442e+09	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	2.422e+09	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	2.432e+09	1.139643e+12

Table 2: Numerical variables mean

	Mean
posX	13.517162
posY	5.896623
posZ	0.000000
orientation	167.162523
signal	-61.703083

Table 3: Mean values for numerical variables

	Mean
posX	13.517162
posY	5.896623
posZ	0.000000
orientation	167.162523
signal	-61.703083

Position

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 removed the posZ variable.

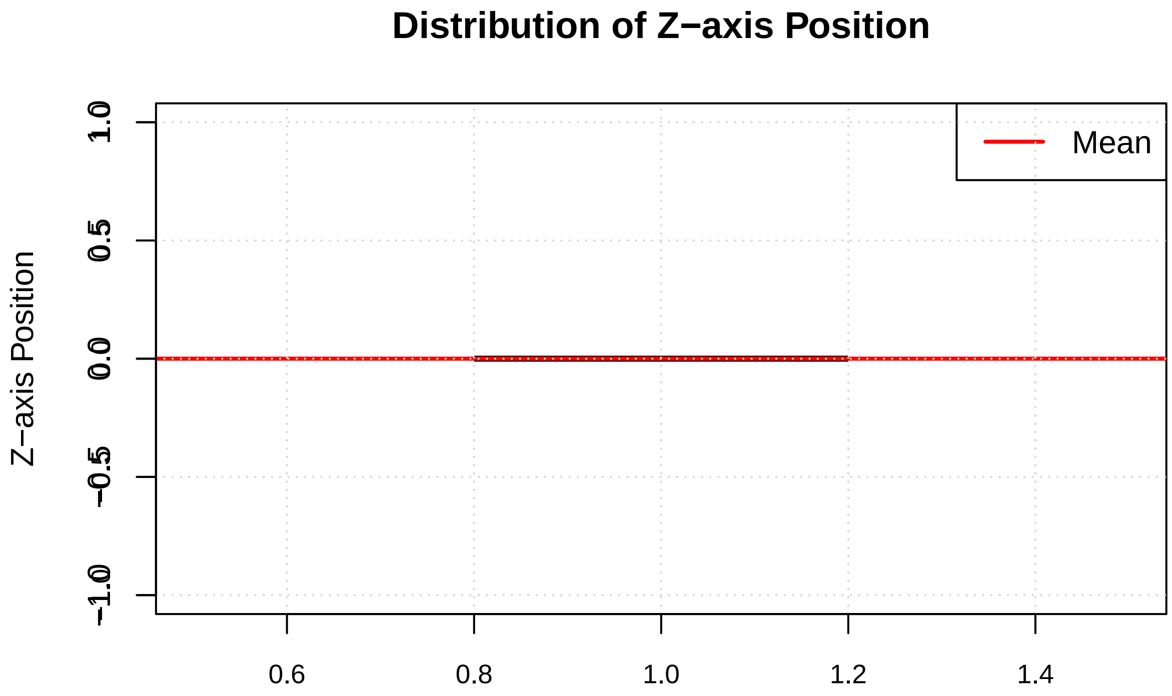
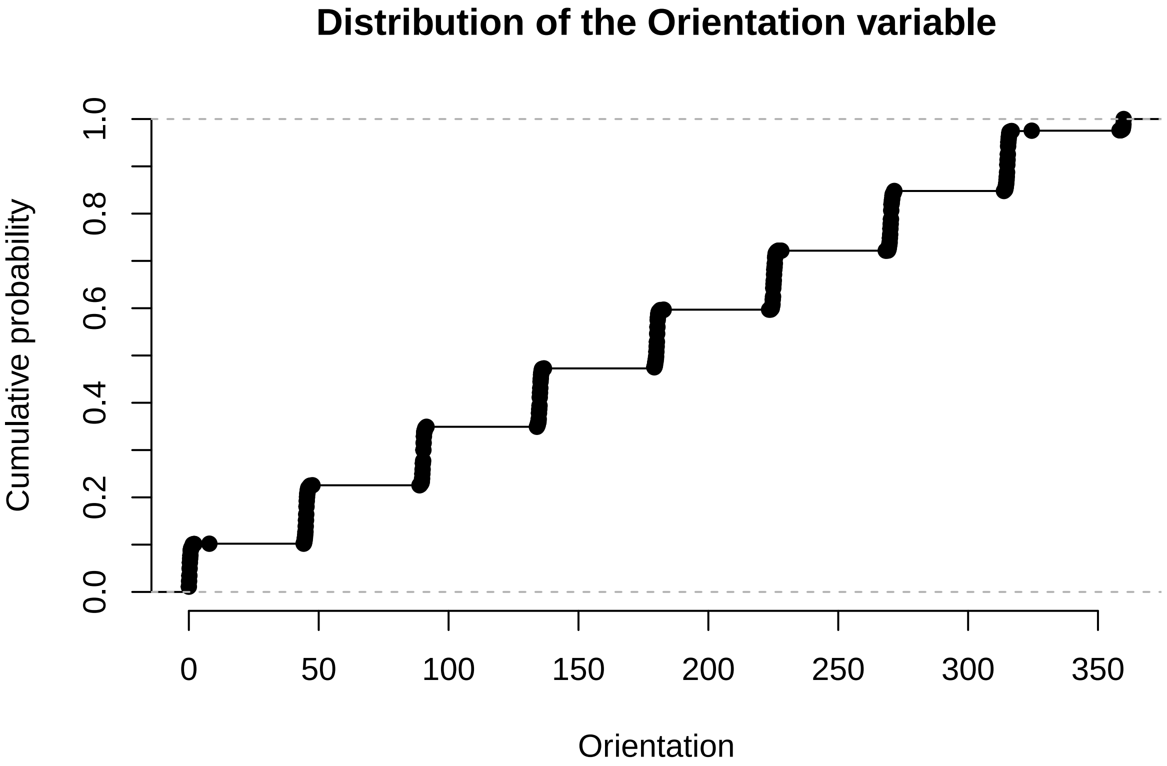


Table 4: Values for posZ

Mean	Max	Min
0	0	0

Orientation of hand-held devices

As provided by the client, the orientation of the hand-held device is supposed to be a set of exactly eight angles from 0 - 315 degrees in increments of 45 degrees. However, in practice, the measured orientations slightly deviate from these eight values. To further examine the distribution of the '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. There is also a value of 360, which should be converted to 0. As instructed by the client, we group the orientation values into bins from 0 - 315 in increments of 45.

Access points

The client provided a list of six access point locations. However, the data collected include 12 access points. This problem arises from the scanning device catching signals from wifi routers on other floors and nearby. We tally the number of observations for each access point and found seven access points with significantly higher number of observations than other access points.

Table 5: Number of observation by MAC address

MAC ID	Observations
00:0f:a3:39:e1:c0	145862
00:0f:a3:39:dd:cd	145619
00:14:bf:b1:97:8a	132962
00:14:bf:3b:c7:c6	126529
00:14:bf:b1:97:90	122315
00:14:bf:b1:97:8d	121325
00:14:bf:b1:97:81	120339
00:0f:a3:39:e0:4b	43508
00:0f:a3:39:e2:10	19162
00:04:0e:5c:23:fc	418
00:30:bd:f8:7f:c5	301
00:e0:63:82:8b:a9	103

Table 6: MAC addresses provided by client

MAC ID
Macs
00:0f:a3:39:e1:c0
00:14:bf:b1:97:8a
00:14:bf:3b:c7:c6
00:14:bf:b1:97:90
00:14:bf:b1:97:8d
00:14:bf:b1:97:81

We confirm the corresponding access points with the data provided by the client on the “Access Point Location” file and filter accordingly.

To confirm the list of access points, we match them with the channel’s frequency. Since each access point coordinates with only one channel, we do not need the `channel` variable for our model.

```
## ‘summarise()‘ has grouped output by ’mac’. You can override using the ‘.groups‘
## argument.
```

```
## # A tibble: 6 x 2
## # Groups:   mac [6]
##   mac                  channel
##   <chr>                  <dbl>
## 1 00:0f:a3:39:e1:c0 2462000000
## 2 00:14:bf:3b:c7:c6 2432000000
## 3 00:14:bf:b1:97:81 2422000000
## 4 00:14:bf:b1:97:8a 2437000000
## 5 00:14:bf:b1:97:8d 2442000000
## 6 00:14:bf:b1:97:90 2427000000
```

4. Challenges

This section discusses challenges we encountered exploring the Indoor Positioning data.

Data formatting

The data the client provided is not in a format ready for analysis. The data file had multiple characteristics that need significant re-formatting:

- It is a text file with minimal formatting.
- The data was not presented in a table format and variables are separated by multiple different separators.
- Each line corresponds to multiple observations and the number of observations on each line is different.

The data formatting process took significant amount of time and required us to learn new functions

Data exploration

- There are different counts between MAC addresses and channels. From the txt file, we realized that there are extra access points that are not included in the testing area.
- We found that there is a one to one relationship between MAC address and channel for the seven devices. We had to delete channel from the 'offline' data set.