MAA

Master Program in Advanced Analytics

Evaluating WI-FI indoor positioning approaches

In a real world environment

Raafat Hantoush

Internship report presented as partial requirement for obtaining the Master's degree in Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

Evaluating WI-FI indoor positioning approaches

In a real world environment

by

Raafat Hantoush

Internship report presented as partial requirement for obtaining the Master's degree in advanced analytics

Advisor: Leonardo Vanneschi

Co Advisor*: Fabian Brosig, Alexander Müller

October 2016

DEDICATION

To the soul of my brother Rodi...

To my Syrian family...

ACKNOWLEDGEMENTS

ABSTRACT

Global positioning system(GPS) does not provide generally a good positioning performance in an indoor location because of many reasons (Henniges, 2012). On the other hand, other alternatives such as the WI-FI technology has become recently in a popular use to provide indoor localization. And that is due to many reasons, such as the wide spread of WI-FI infrastructure in the indoor environments and the low cost of this technology. This study attempts to evaluate different WI-FI indoor positioning approaches in a real world environment. In particular, in retail stores and shopping malls. The pros and cons of each one of these approaches are pointed out.

The main purpose of this study from the company perspective is to explore the state of the art methods and the cutting edge technologies of the WI-FI IPS and to come up with an improvement of their indoor localization system. This system forms the core of the company`s retail-analytics product that uses a Wi-Fi positioning technology to provide indoor location based services for the customers and helps retailers to better understanding their businesses.

KEYWORDS

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

802.11   IEEE standard to implement wireless local area network.

AP   WI-FI access point.

RSS   Received signal strength.

NTP   Network transfer protocol**.**

GPS   Global positioning system.

RF   Radio frequency.

The company   Minodes company.

Mapper   The system that is used to estimate the clients` location.

MIMO   Multiple input multiple output.

IPS   Indoor positioning system.

LBS   Location based services.

1. INTRODUCTION

This document represents the whole study that was done for the purpose of evaluating different approaches for indoor localization using WI-FI Technology. This study was held in Berlin, Germany in cooperation with Minodes company. The role of the author (Raafat Hantoush), as a research data scientist, was to conduct a practical scientific research to improve their retail analytics product. Specifically, to improve the company`s currently used WI-FI indoor positioning system by exploring the state of the art and cutting edge technologies in the indoor positioning approaches.

Firstly, this document starts with giving the reader an overview of 802.11 standard which represents the main medium of providing the indoor positioning services. That includes describing the building block of this standard and the Frames structures. It also gives an overview of the indoor positioning system. Then, this document points out the advance in this field and the state of the art of different WI-FI based methods with mentioning in detail the advantages and disadvantages of each one of these approaches comparing with the other approaches.

Secondly, a brief about the company`s retail analytics product is introduced. That includes mentioning the product features and the real problem that this product solves. Then, a detailed explanation is introduced to show how the company`s current indoor positioning system works.

Thirdly, the experimental setup is explained in detail by describing the field, locations, resources and other conditions that were applied during the experiments. Then, a detailed exploratory data analysis is presented for better understanding the underlying relationships and the analysis of the association between different factors that affect significantly the performance of the positioning system. After that, a data preprocessing and the feature engineering are presented. Also, an explanation of the conducted implementation of both of the modeling based and cell-of-origin approaches are presented.

Finally, the results and the conclusion is introduced. Also, the current limitation and obstacles that faced this research are described in detail. And the recommendations for the future work are also pointed out.

2. LITERATURE REVIEW

2.1. FOUNDATIONS

2.1.1. What is 802.11?

WI-FI is a wireless technology that uses a radio waves to connect computers and other network-enabled devices within a local area networks and connect them also to the worldwide networks. WI-FI technology uses the family of common standard IEEE 802.11 that is defined and managed by Institute of Electrical and Electronics Engineers (IEEE). The base version 802.11-1997 was released in 1997 and it was the first wireless networking standard in the family. Other standards in the same family followed the base one and they included several improvements and extensions in the specifications and corrections for the previous problems. Most of the conventional IEEE 802.11 standards use 2.4 GHZ and 5 GHZ bands (Gast, 2002, Chapter 1).

2.1.2. 802.11 networks building blocks:
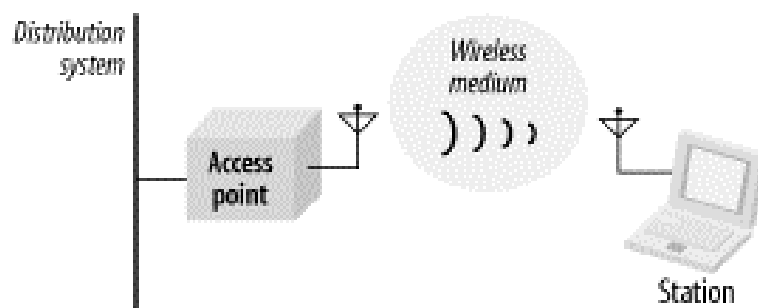
Figure 2.1-WI-FI building block , [Gast ,2002]

1. Distribution system is the backbone medium that transmits the frames between the access points to stay connected with each other to exchange information about the mobile stations. Ethernet is the most popular network technology that have been used to implement the distribution system.
2. Access point is the networking hardware that is used to connect wireless devices to a wired network as the main function in additions to other functions.
3. Wireless medium: 802.11 standard allows that many physical layers can be used as a medium to transfer frames between stations but the radio frequency (RF) physical layers proves to be the most popular ones.
4. Stations: any Wi-Fi-equipped computing devices can connect to a wireless network through a wireless network interface card and can send and receive data to other devices on the network.

2.1.3. Infrastructure wireless network design

The access point and the stations that are reached by it forms an area called basic service set (BSS) that is identified by 6-byte identifier called BSSID. In Infrastructure mode wireless

network type, when a client station within the BSS area wants to communicate with other stations, it should send the frame first to the access point and then the access point will forward the frame to the destination station. Using the access point (AP) is the main characteristics of this type of network (infrastructure mode).
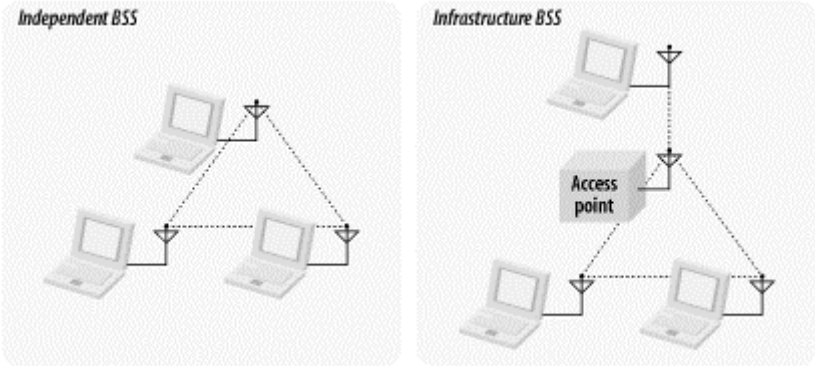
Figure 2.2 - Network infrastructure mode, [Gast,2002]

Extended service set (ESS) is formed when several access points connect to each other (and hence their BSS`) via a distribution system that could be a wired or wireless.
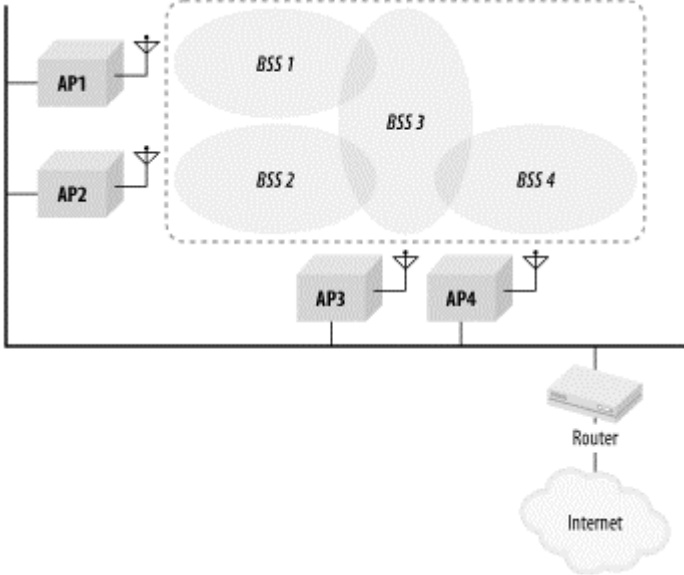
Figure 2.3 - Extended service set, [Gast,2002]

2.1.4. Wi-Fi frames types

Comparing to Ethernet frames, Wi-Fi frames is a little bit more complicated because it needs more management and control features to be involved. Three main types of frames are existed; Data, control and management frames (Gast, 2002, Chapter 4) .

- Data frames are the frames to carry the data from station to another which is the main purpose of the network is transfer data.

Figure 2.4 - WI-FI generic data frame, [Gast,2002]

- Control frames: these frames are used to control and manage the access to the wireless medium. Some of these frames are:

  - Request to send frame (RTS): this control frame is used to request access to the wireless medium to transfer frames. This frame type has no frame body and the frame header is followed directly by FCS field.

  - Clear to Send (CTS): it is the answer of the previous request frame and it is contains only the frame header as RTS frame. It is sent by the receiver of the RTS frame to tell if the transmitter can start to send the large data frames over the wireless medium.

  - Acknowledgment (ACK): it provides a positive acknowledgment and is used with any data transmission, including plain transmission.

- Management Frames

As mentioned before, wireless networks need management features more than the wired networks. 802.11 standards categorize this features in three processes; choosing the appropriate wireless network, authenticating the station to grant permission to connect to the network and, finally, associating the client into the network.

There are several types of management frames. We will mention the most important ones for our study:

  - Beacons: In infrastructure mode WLAN networks, beacon frames are transmitted by access points on a regular basis to identify the existence of the network. The area in which the stations are able to receive the beacons forms a BSS area.

- Probe Request frame: This type of frame is transmitted by the client to scan for 802.11 networks in the area. The identifier of requested BSS network (SSID) along with the supported rates that the client station can communicate with, are transmitted in the frame body of this request to the access points. A broadcast SSID is used in case there is no specified or already defined network.
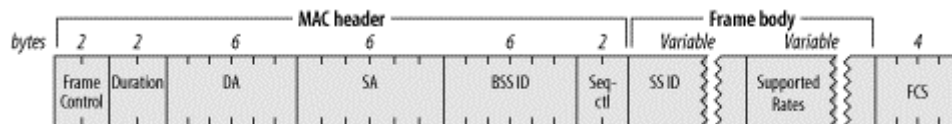
Figure 2.5 - WI-FI probe request frame, [Gast,2002]

Figure 2.6 - WI-FI probe request detail, [Wireshark]

- Probe response frame: This type of frame is sent by the access point (in infrastructure networks) to respond to the incoming probe request. The format of this frame is similar to the beacon frame and hold all the beacon parameters that let the client station to join the network later.
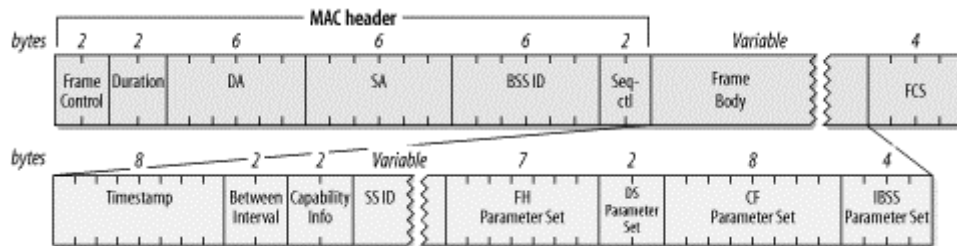
Figure 2.7- WI-FI probe response frame, [Gast,2002]

- Authentication request, authentication response are used for authentication process between the client and the access point. Association request and association response are the management frames that associate the client to the network.
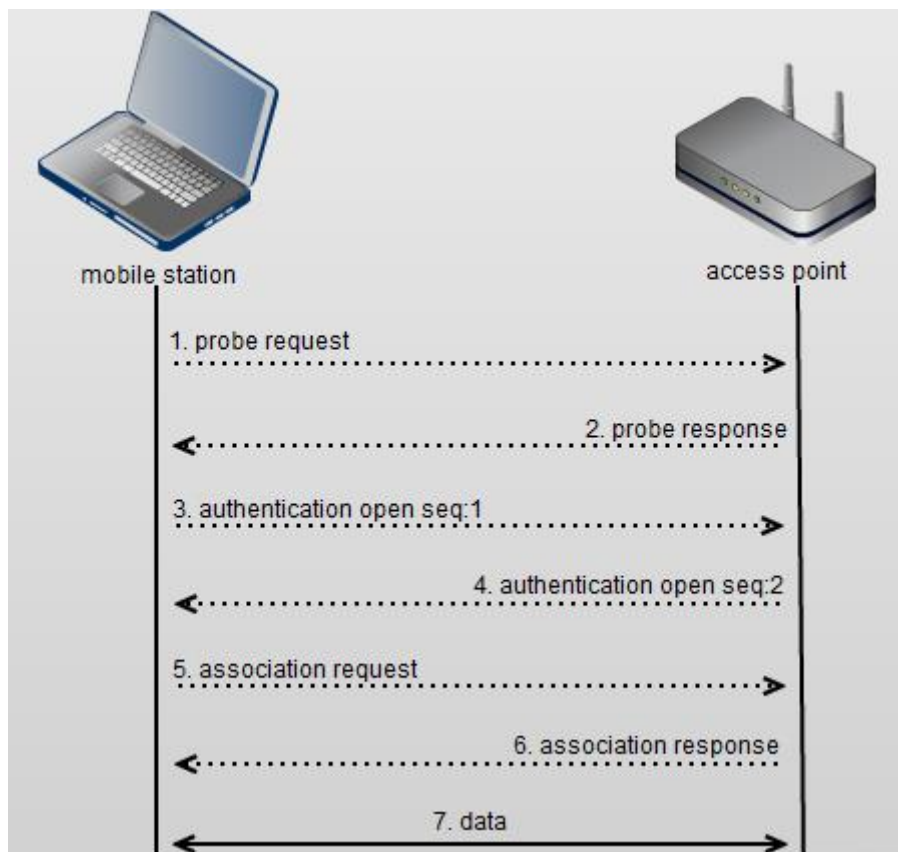
Figure 2.8 - WI-FI management frames, [Cisco]

2.1.5. Location based services (LBS)

It is defined as services that combine the geographical position of the mobile device with other information to provide an added value to the user. These kind of services started to be

in use in military applications by the US government in 1970s through a satellite system which provides a global positioning service for people and objects (El-Rabbany, 2002, Chapter 1). This system is called GPS. In 1980s the US decided to democratise this system followed by Europe that built a similar system to provide this service. Since then other industries started to use this service to provide a better product and service with using the user`s information. Location based services became widely and heavily used especially when the smartphones become a main stream in the market. A good example automotive industry starts to embed this system in the cars to provide a navigation service. Other popular use is providing a marketing and advertising services inside the shopping malls and stores based on the user`s location data.

2.1.6. Indoor positioning system (IPS)

GPS system is used to be an efficient technology in providing an accurate positioning service in the outdoor locations. In contrast, this service suffers some obstacles that affect negatively this accuracy, such as signal attenuation caused by construction materials. Other reason is the multipath loss due to many reflections at surfaces (Ozsoy, Bozkurt, & Tekin, 2013). For these reasons, other technologies like WI-FI technology has been proposed to be the alternative of GPS in indoor positioning system and it shows a better positioning accuracy. Bluetooth, RFID are other examples of wireless technologies that can be used for indoor positioning system and they give good results as well. It is also an important to notice that Radio waves (RF) based positioning systems are not the only alternatives of GPS. This study will focus only on the WI-FI based indoor positioning systems (WPS) (Curran et al., 2011, pp. 61–78).

2.2. STATE OF THE ART

There are many approaches for locating client devices in an indoor environment using WI-FI Technologies. These approaches can vary by several factors such as the hardware needs, performance, accuracy, cost and maintenance. It is possible and useful to categorize these approaches based on the mechanism each of these approaches uses to estimate the location in the indoor environments. A taxonomy of WI-FI existing positioning approaches is:

- Proximity Based Method:
    - Cell of Origin (Cell Identity).
- Triangulation Based method:
    - Angle of Arrival (AoA).
- Lateration (distance) based:
    - Time of Arrival(TOA) or Time of Flight(TOF).
    - Time Difference of Arrival (TDoA).
    - RSSI Based method.
- Pattern recognition based method:
    - Fingerprinting.

It is woth mentioning that sometimes the time of flight (time of arrival), time difference of arrival methods are considered as a separate time based methods. But in this study it was preferably categorized as a distance-based method because the calculation of the location at the end is based on the distance that was concluded from the time measurements. In the following we

2.2.1. Proximity Detection (Cell-Of-Origin)

This is the simplest approach for positioning. The position of the client station or user device is determined by the nearest cell of origin (Access point) (Henniges, 2012). The nearest cell is identified by the strongest signal that the client has received from the cell(node) within its range. This means that if the client station is within a range of more than one cell Access Point, the cell with the strongest signal will give its position to the client station. The Cell of origin approach is not only used with WI-FI based techniques but with other wireless technologies as well like GSM (Cell –ID) where it is more popular. (Farid, Nordin, & Ismail, 2013).

This approach is simple, fast, real-time positioning and no need for any hardware modification. Another advantage of this approach is the low impact of multi-path effect on the perofrmance of the approach. On the other hand, the serious limitation of this approach is the low and the insignificant accuracy of the location estimation. Because it reports the location as within the range of the nearest cell (node). Also, as another limitation, it suffers from the mutli-floor problem when the store is a multi-story and unfortunately this is the case of the majority of real world stores/shopes. And it is quite difficult using this approach to tell which floor the visitor is in and that is due to the signal overlap between stories.
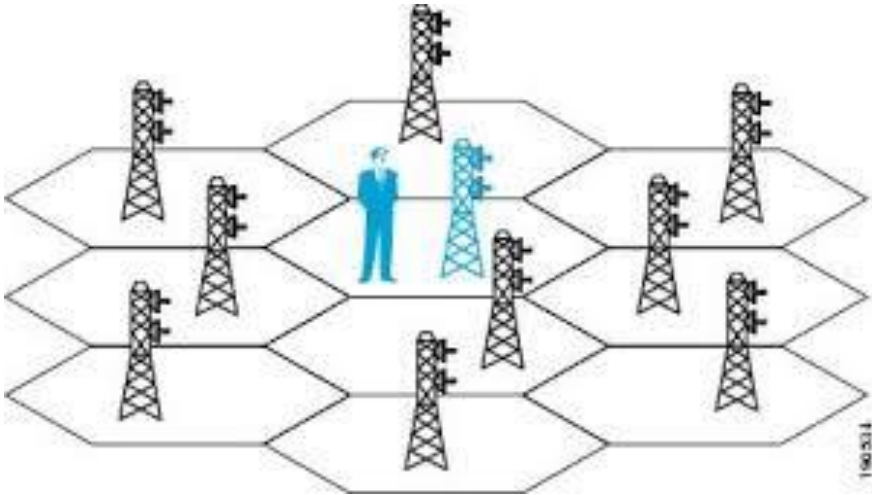
Figure 2.9 - Cell Of Origin, [Cisco]

2.2.2. Triangulation Based (Angel Based Methods - AOA)

In this method, the positioning of a client station determines relatively to the position of two known access points using geometric calculation called triangulation. The distance is calculated based on two known angles and the fixed distance between the two stations (See the figure below). This method requires the stations to be equipped by special antennas to be able to detect the angles of the incoming signals (Henniges, 2012). A directional antenna or a multiple antenna array (MIMO WI-FI interfaces) can be used for this purpose.

This technique provides a better accuracy than other techniques, specifically in direct line of sight enviorments. Also, it provides a real time positioning.

This technique has several drawbacks and limitations. The modification of hardware is needed to be able to perform the calculations for this approach. The stations should be equipped with special antennas(directional antennas) or multiple array antennas such as MIMO-Enabled WI-FI interface cards to be able to determine the angle. Because of this modification of hardware, the approach can be seen as costy solution and impractical, especially in the big stores and shops where many stations are needed to cover the whole indoor areas.

Another limitation of this approach is that its accuracy is affected by signal reflections.
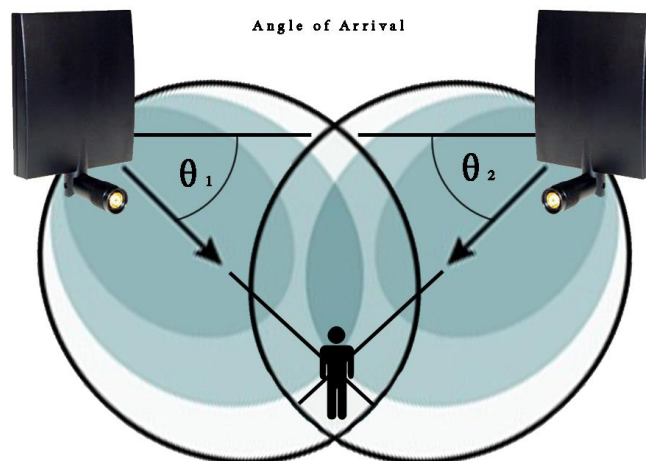
Figure 2.10 - Angle of arrival, [Henniges, 2012]

2.2.3. Lateration (Distance) Based

In this method, the position is determined based on the distance measurements. In trilateration technique, the position is calculated based on the distance from three reference points(stations) (Henniges, 2012). The distance from the station to the mobile device is used as a radius of the circle that its circumference forms the potential position of the client device. The distance from the second station forms the radius of the second circle and the two intersection points of the two circles are the potential location of the device. So three circles needed to determine the exact position and a fourth station needed to give the position in three dimensional environment like multi floor building. As it is clear that the localisations in the lateration based technique depends a lot on the calculation of the distance. A three methods to calculate the distance are already in use and described below.

2.2.3.1.   Time of arrival (TOA) or time of flight (TOF)

It uses the time measurement from the transmitter (client device) to the receiver (station) to calculate the distance between the client and the station (see the figure below). Since the calculation depends on the difference between the time of departure and time of arrival, all the receivers and the clients should be highly synchronized (Headquarters, 2008, Chapter 2). Any difference in the clocks affects negatively the distance calculation and hence the positioning estimation perofrmance.
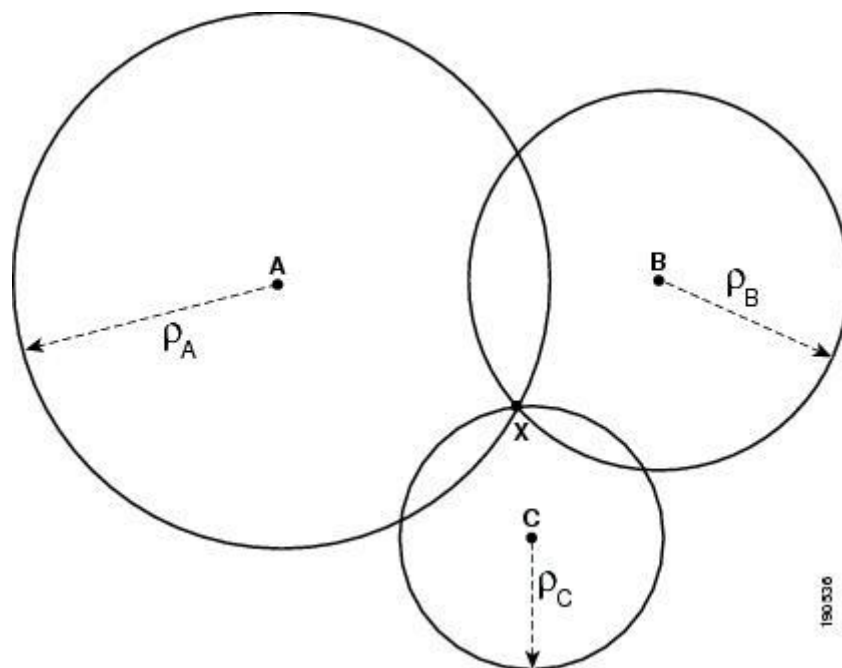
Figure 2.11 - Time of arrival (TOA) , [Cisco]

It is clearly seen that the one critical drawback of this approach is the critical need of a very accurate time synchronization between the client devices and the stations in order to perform a correct calculations. Given that a very small difference of time produces a large difference(error) in location estimation. Another drawback of this approach is that in a real world scenario, more likely the client devices are not under control or, in another meaning, there is no access to these client devices in order to be able to perform the calculation and the synchronization.

It is worth mentioning that TOA approach provide better performace and accuracy in the location estimations in more open, large and high buildings where the multipath problem could be at the minimum.

2.2.3.2.   Time difference of arrival (TDoA)

It is similar to TOA in terms of using the time of travel to calculate the distance. But instead of using the travel time from the client to the receiver, the difference in time arrival between each receiver are used to find the distance between each sensor. Each difference-of-arrival time measurements forms a hyperbolic curve on which the client lies on (see the figure below). The intersection between these curves is the position estimate of the mobile device (Headquarters, 2008, Chapter 2). In contrast to TOA, There is no need for synchronization between the client and the stations. However, the clocks of the stations should be synchronised in this method.
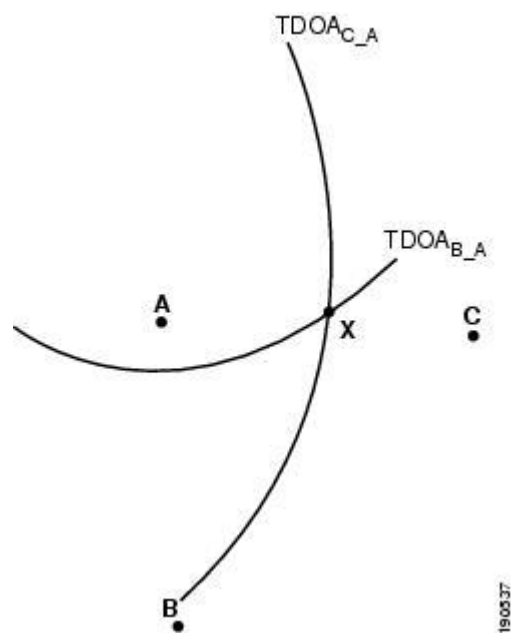
Figure 2.12 - Time difference of arrival, [Cisco]

One advantage of this approach is the ability to perform a real time positioning. Also, For the previously mentioned reasons, it is clear that this approach is more popular than the previous one. Because it is easier to to implement and no need for the synchronisation with the client devices.

But on the other hand, it is still dificult to assure the highly accurate synchronizations between the stations (receivers). And it is similar to TOA in terms of its effect by the multipath and interference that could happen in indoor environment.

2.2.3.3.   Received signal strength (RSS)

This method uses the attenuation of the emitted signal to calculate the distance of the client device from the reference node (Henniges, 2012). The received signal strength indicator (RSSI) is the most widely used signal related feature and it is measured by decibels.
One advantage of this approach is the significant cost effeciency of the implementation on a large scale. Because there is no need for any hardware modification or additional sources. Another advatnage which makes this approach similar to the previous techniques is the ability to perform real time positioning.

On the other hand, this approach shows a poor performance in real world indoor environment because its high effect by the multi-path and signal reflections.

2.2.4. Pattern recognition based method (fingerprinting)

This method uses pre-built database that contains a signal patterns (fingerprints) from the area of the interest that has been taken previously in a phase called a training phase (Navarro, Peuker, Quan, Clark, & Jipson, 2010). Each fingerprint contains all the information about the stations and the received signal strength(RSS) values. A location of new client device in the recorded area can be determined based on a similar pattern already existed in the fingerprint database. The similarity between the new fingerprint and the training data can be found by applying several machine learning and pattern recognition techniques.

The main disadvantage of this approach is that it is a time consuming and it needs an operational effort comparing to the previous methods. That is due to the need of a cailbration phase to collect the fingerpint database (training set). Also, because of this drawback, this method could be seen as non-real-time positioning approach. Although it provides online location estimation using the pre-built pattern based model. Another serious drawback of this technique is that in any modification in the indoor environment, makes the fingerprint database outdated and affects negatively the accuracy of the positioning system over time. Hence, it is required to rebuild the

database from the beginning. This serious problem is not even existed in the previous mentioned approaches because no historical states are stored.

On the contrary, the accuracy of RSSI fingerprinting based method is remarkably good. Also, as a very positive point is that no need for hardware modification or any special high-cost additional hardware. That makes it a very popular approach in the commercial market due to the cost effectiency on the large scale and due to the very practical and good localisations estimations that provided by this technique.

Not surprisengly, that the company uses the RSSI fingerprinting positioning approach as the core and the only approach for its retail analytics product. In this study, as shown later, a lot of experiments and deep analysis were applied mainly on the RSS fingerprinting approach to try to improve the company currently used approach, and even trying include other approaches with this approach.

3. THE COMPANY `S STATUS QUO

3.1. COMPANY BASIC INFORMATION

- Name: Minodes GmbH
- Company logo: MiNODES
- Physical Address: Friedrichstraße 224, 10969 Berlin, Germany
- Web address: www.Minodes.com
- Email: info@minodes.com
- Telephone: +49 30 208 473 870
- The responsible from the company side: Dr. Fabian Brosig, Alexander Müller
- Email of the responsible: fabian.brosig@minodes.com

3.2. COMPANY PRODUCT DESCRIPTION

The company provides a retail analytics cloud-based platform that helps the offline retailers to get better understanding of their customers' behaviours and also their sales and marketing operations.

The insights and reports are published online to the retail owners. These insights answer a questions such as (see the three following figures):

- How many visitors approximately visited a specific zone (department) in specific time in specific store?
- How the customers visiting path look like inside the store?
- How much time the customers spend looking to the new collection of cosmetic products that was promoted in Zone A in the first floor in Store B?
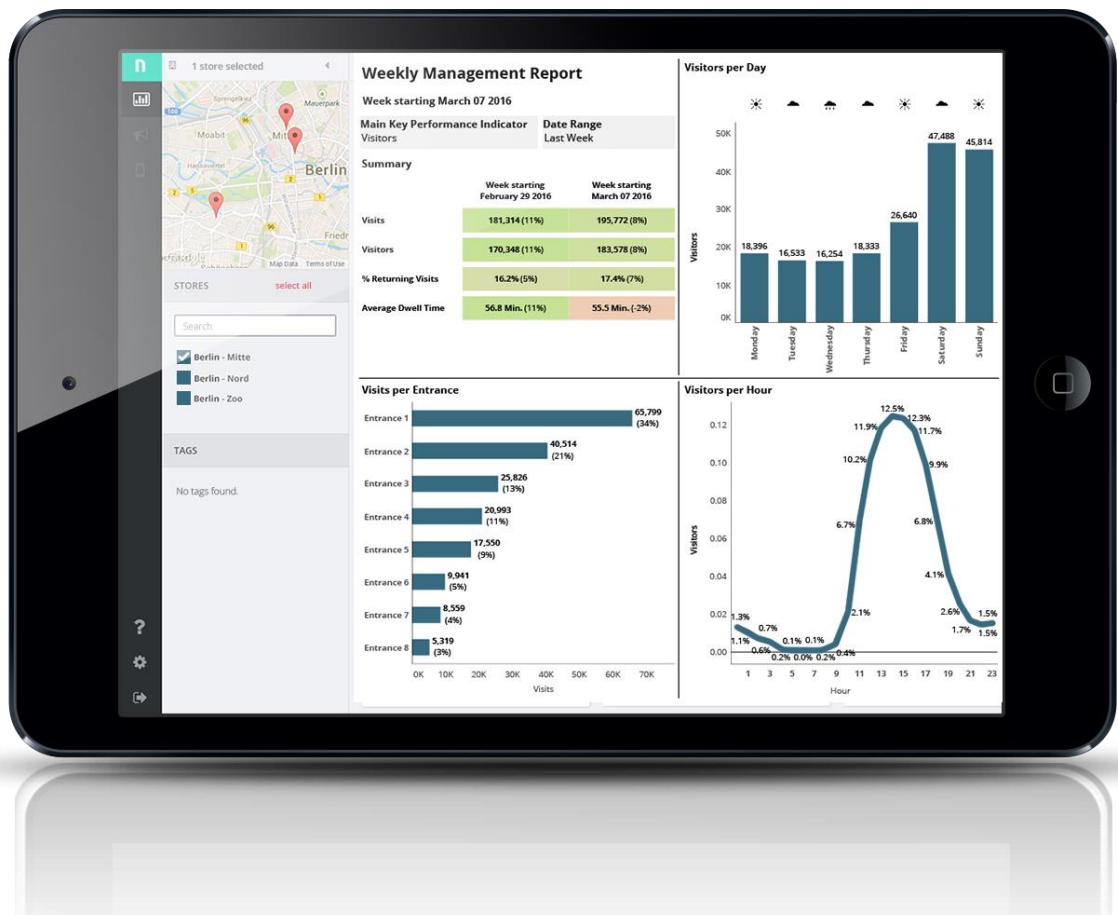
Figure 3.1 - The company`s product main dashboard, [Minodes]

Bereichsmigration

Zeitraum	Bereich	Top Pfade
Letzte Woche	Adler	10

Bereichsmigration von Adler zu...
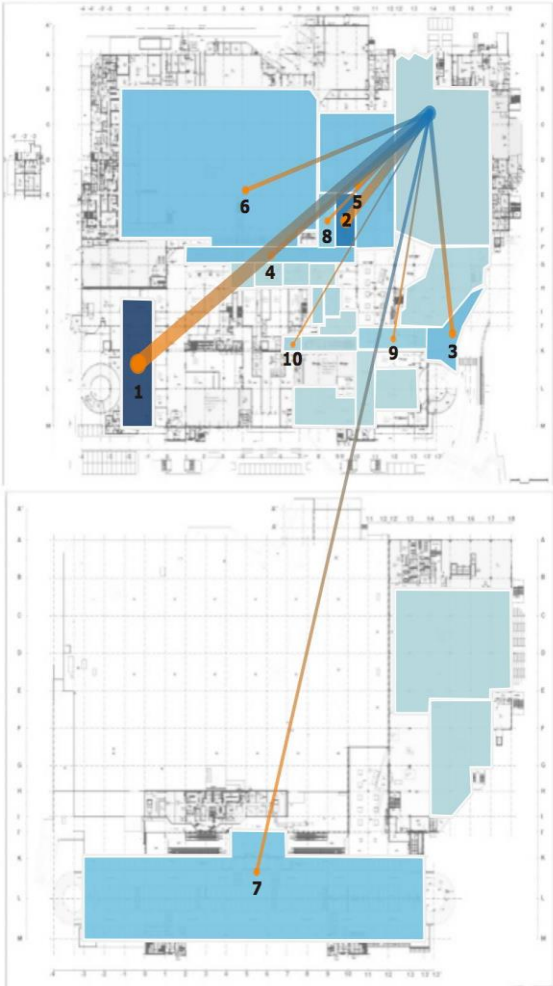
Figure 3.2 - The company heatmap report 1, [Minodes]

Besuchsverlaufsanalyse

30. Juli 2016 - 30. August 2016

Top 10 Besuchspfade
Klicken Sie auf einen Pfad oder einen Bereich um den Verlauf hervorzuheben

Zeitraum
Letzter Monat
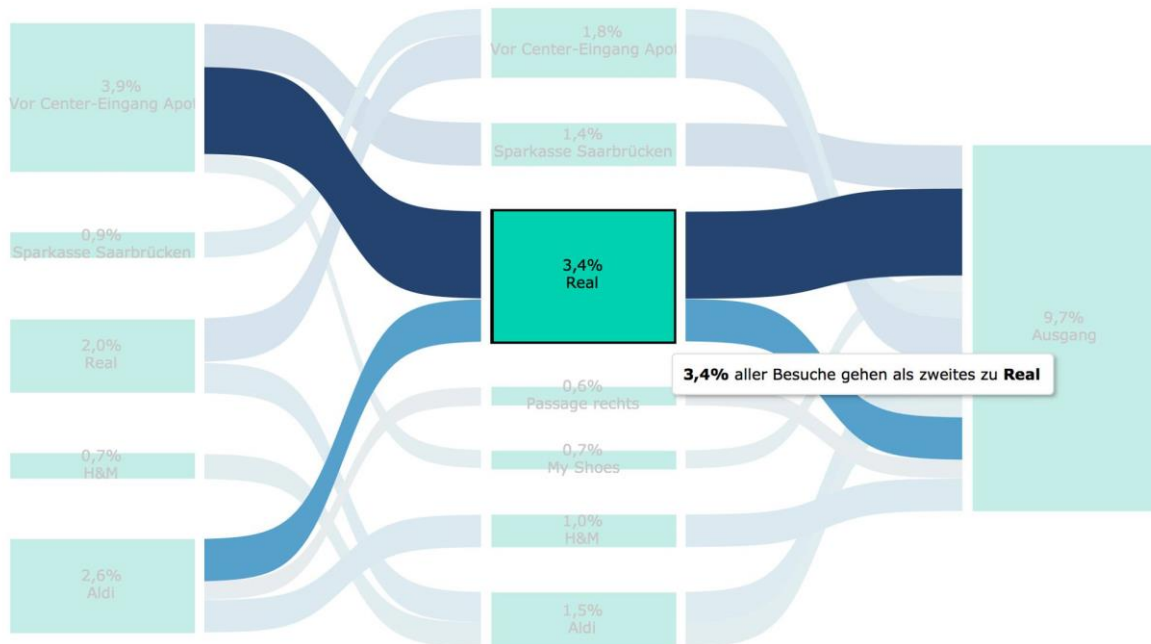
Figure 3.3 - The company`s product heatmap Report 2, [Minodes]

3.3. HOW IT WORKS

The company does that via installing nodes (WI-FI access points) in the retailers' stores and shopping malls. These nodes (APs) listen to the WI-FI signals (probe requests) that are broadcasted by the visitors (clients) devices. The incoming signals information will be sent to the company central servers over the internet from all the nodes that are installed in the stores. These data will be analysed by the cloud-based company online software tools and eventually the clients locations in a specific time will be estimated using machine learning techniques.

Figure 3.4 - WI-FI access point, [TP-Link]

A mesh network was defined among all the nodes within each store to insure a stable communication medium to transfer data between nodes and to the internet via the master node.
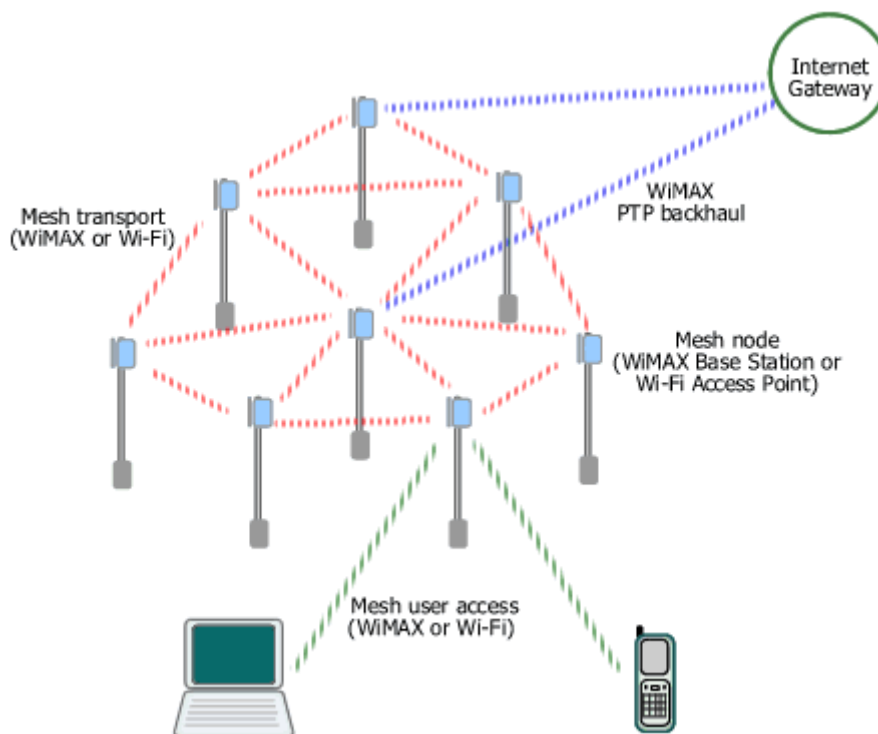
Figure 3.5 - WI-FI mesh network, [conniq.com]

The company uses currently in-house developed machine learning system to estimate the clients locations.The company call it a mapper. This system depends on a training set to be able to learn the model. The operation team is responsible for collecting the training set data via a calibration phase. This phase is very important step and it should be done for each store of the company clients. The calibration phase happens after the setup of the

company`s system in every new store. The whole setup steps described in breif in the following:

- The zones should be planned in advance before installing the nodes in the store. the zones should be drawn on the floor plan and should follow some conditions. For instance, the zone should be approximately 40 square meters.
- The neighbouring area outside the store, especially in the very busy shopping street, should be also planned into zones if it possible.
- After planning the zones carefully and drawn them on the store floors plans, the installation of the nodes in specific store can start and should take in consideration some concerns like the zone size and locations.
- The nodes (WI-FI APs) should be installed and ready for receiving the incoming signals. Also the Mesh network should be configured properly and the master node should be connected to the internet to be able to send the data to the central servers regularly and frequently.

3.3.1. Calibration phase (training phase)

The calibration phase starts after making sure the setup was done properly.

- The operation team start the calibrating phase by walking in constant speed in every square meter in every planned zone in the store. they do that by holding the measurement phones set that send WI-FI pings frequently. This measurement phones set include four to five phones from the most popular brands (Samsung, Sony, IPhone, HTC, Nexus). And to be able to reference every square meter during the calibration for better results, the "WI-FI Analyser" mobile application is used to increase the frequency of sending the pings during the calibration.

- In the same time of the calibration phase, an in-house developed application(Realytics app) is used to label the reference data. Labelling the data means defining the zones that pings came from based on prior knowledge of the phone device location at a specific period of time.
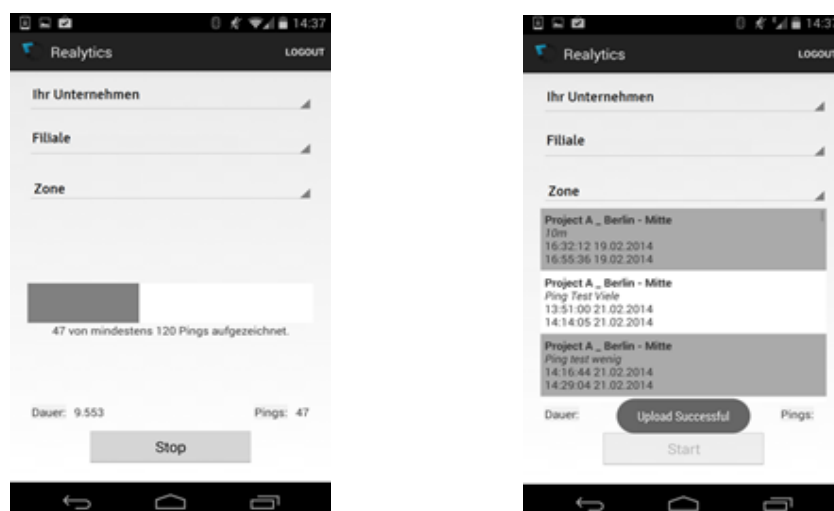
Figure 3.6 - Realytics app , [Minodes]

4. METHODOLOGY

4.1.1. Experiments locations

The experiments and evaluation of the Wi-Fi positioning approaches, as well as all kinds of data analysis for the Wi-Fi data, have been located in the following locations:

- The company building: most of the experiments have been conducted in the company floor where all the experiments conditions were fully controlled and manageable.
- Neighbours locations: other experiments, especially those related to measuring some factors that affect the data quality and the analysis of the data,were located in the two floors above and down the company floor in the same building.
- customer's sites:For the data analysis purposes for some experiments, a live data was collected from the customer's sites for a specific period of time from the sensors data. These customer's sites such as Kadewe store, Kamden Market and ECC.

4.1.2. Experiments components

- The measurement set: several mobile phones have been used for conducting the fully controlled experiments. These phones have these conditions:
  - WI-FI-enabled phones: the WI-FI of the mobile phones has to be enabled and switched on during the whole period of the experiment.

Figure 4.1 - Measurement set phones

- The battery of these phones should be charged enough to avoid switching off the phones during the experiment time. (it happens that during some experiments, some phones were switched off after the starting time of the experiment and that results in excluding this phone from the whole experiment or repeating the experiment another time with the whole measurement set.
- The application WIFI-analyser should be up and running during the whole time period of some experiments to increase the frequency of generating Wi-Fi signals (pings) which results in generating more data for analysis purposes. However, some experiments were held without using the Wi-Fi analyser to increase the frequency such as ping delays analysis.

Figure 4.2 - WIFI Analyzer

- WI-FI access point: one or more Wi-Fi access points have been used to collect the Wi-Fi signals from the mobile phones and send this data to a central server for processing it. An access point (node) that is packaged and supports 802.11 standard with support of two networks bandwidth 2.4 and 5 GHz. (See the following figure).

Figure 4.3 - WI-FI access point(node) , [TP-Link]

- A mesh network connecting all the access points (nodes) with each other to synchronise the time and to transfer the data.
- Internet enabled Access point acts as a master node in the mesh network. Its main function is to receive all the data from the other nodes and transfer it periodically to the central server to do the calculation.
- Portable node for some experiments where one node or more have been plugged with portable power charger to enable the free movement of the nodes during the experiment.
- Central server to store all the data and do the processing. (the company server has been used for this purpose without using any dedicated server for this study experiments.

4.1.3. Experimental field

The company office has been prepared to be mainly used for most of the experiments. Dedicated nodes have been located in different locations in the office to cover the predefined zones. The setup has been done to assure that each zone was covered by one node. After that, a mesh network was configured properly between these nodes (Aps). The follwing table shows the IDs of the nodes that have been used to cover the zones that were named by the department name that is located in each of these zones. Also, the following figure shows the floor plan of the company marked with the nodes on top of it.

Table 4-1 - Minodes floor nodes installation

Store id	Node Id	Zone ID	Mac Address
23	13	Kitchen	E8DE27B8B9CA
23	23	Kitchen Table	E8DE27E12Bec
23	12	HR Office	E8DE27E12A4E

23		BizDev	Master
23	8	Near entrance	E8DE27B8B352
23	9	Management	E8DE27C13D64
23	7	development	E8DE27E121CA
23	6	Library	E8DE27E139E2
23	1	Meeting room	E8DE27E13CB4
23	5	Sales	E8DE2769BCAC
23	4	Operation	E8DE27E139E6
23	2	Product Management	E8DE27E13AB2
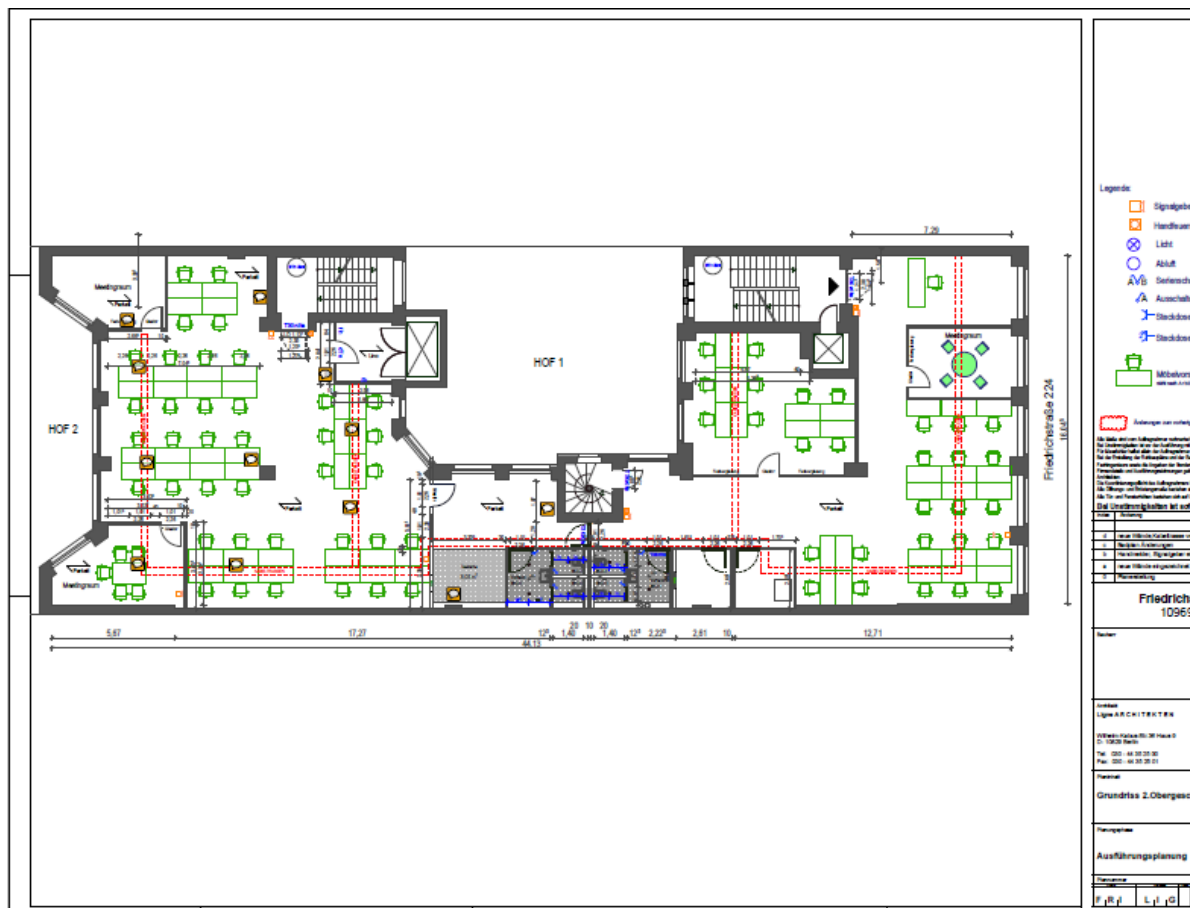
Figure 4.4 - The experimental field, [Minodes]

4.2. EXPLORATORY DATA ANALYSIS

The purpose of this analysis is to understand deeply the relationship between different variables and the general effect on the accuracy of the indoor positioning system. This analysis focused mainly on the signal strength (RSS) value because it is the most important predicror(factor) for the majority of the indoor positioning approaches. Also in this analysis it was tried to answer the most important questions that was raised by the company and the curiosity of the researcher. For answering each of these quesions, multiple test scienarios were applied with raising the strength and weakness of each one of these tests.

The following figure shows the data format that had been used in all the experiments. This data was captured from a specific node(AP) within a specific period of time.

Out[19]:

	phone_mac_address	phone_vendor	signal_strength	timestamp
0	b9898e1e0f7b7fcdf08685581a14a565e7601cd4fc87d6...	d4:20:6d	-61	1458303960.269868
1	b9898e1e0f7b7fcdf08685581a14a565e7601cd4fc87d6...	d4:20:6d	-59	1458303960.289906
2	b9898e1e0f7b7fcdf08685581a14a565e7601cd4fc87d6...	d4:20:6d	-63	1458303960.315588
3	b9898e1e0f7b7fcdf08685581a14a565e7601cd4fc87d6...	d4:20:6d	-66	1458303960.339826
4	b9898e1e0f7b7fcdf08685581a14a565e7601cd4fc87d6...	d4:20:6d	-65	1458303961.291948
5	b9898e1e0f7b7fcdf08685581a14a565e7601cd4fc87d6...	d4:20:6d	-59	1458303961.363896
6	b9898e1e0f7b7fcdf08685581a14a565e7601cd4fc87d6...	d4:20:6d	-67	1458303961.411164

Figure 4.5 - sample data

4.2.1. Q1. What is the highest possible signal strength value that sensor antenna can emit?

- Findings:

    - It seems that the highest value depends on the phone itself. For e.g, Nexus and HTC gave the highest possible value -2,-3 respectively whereas Iphone and Sony xperia gave the lowest possible values -14,-17 but still all of them in -10s range.

    - HTC and Nexus gave the extreme lowest signal values (the rare signals) and it seems it is correlate with the ping frequency, the more the device pings the stranger values.

- Insight and potential use:

It could be useful to use this insight for data pre-processing purposes and outlier detection.

- Test cases:

  - Putting the measurement phones exactly near and in touch with management`s node.

  - Putting the measurement phones exactly near and in touch with development`s node.

- Test weaknesses:

  - The phones were put exactly near the sensors but probably the sensor antenna location and direction was near to some of the phones more than others.

  - The test was applied only for 5 different phones.

- Test strengths:

  - Applying the test to two different sensors.

Table 4-2 - Highest signal strength

	Node#1	Node#2
Phone_name	Signal_strength	Signal_strength
other_rafael_sony	-17	-13
yellow_htc	-2	-8
yellow_iphone	-14	-16
yellow_nexus	-3	0
yellow_sony	-7	0

4.2.2. Q2. Is there any device offset effect and how it affects the data quality?

- Findings:

  - The signal strength values ranges and variations are depending on the phone vendor. Nexus and HTC phones recorded the highest variation whereas Iphone recorded the lowest (50 units, 30 units respectively).

- HTC and Nexus recorded the highest values comparing to Iphone who emitted the lowest RSS values.

- The phones from the same vendor show a similar behaviour which makes sense and gives us a positive advantage to categorize the phones by vendor regardless the phone model. e.g. two sony phones.

- The ranges.

- Insights and potential use:

  - There is a device offset and could be useful to take it in consideration in the data preprocessing phase especially for the model based approach.

  - The similarity of vendor`s phones behaviours could be helpful in categorizing the phones by vendor regardless the model numbers.

- Test cases:

  - Putting the phones near each other in the same place for the same period of time and measure the signal strengths from a specific node.

  - Putting the phones near each other and changing their locations over time for five different locations and measuring the signal strengths from a specific node.

- Test weaknesses:

  - The phones were put exactly near each other but probably the sensor antenna location and direction was near to some of the phones more than others.

  - The test was applied only for 5 different phones.

- Test stregnths:

  - The test repeated three times and showed the same results.

  - The test was applied in real world environment.

- **Test case #1(Phones in the same place)**
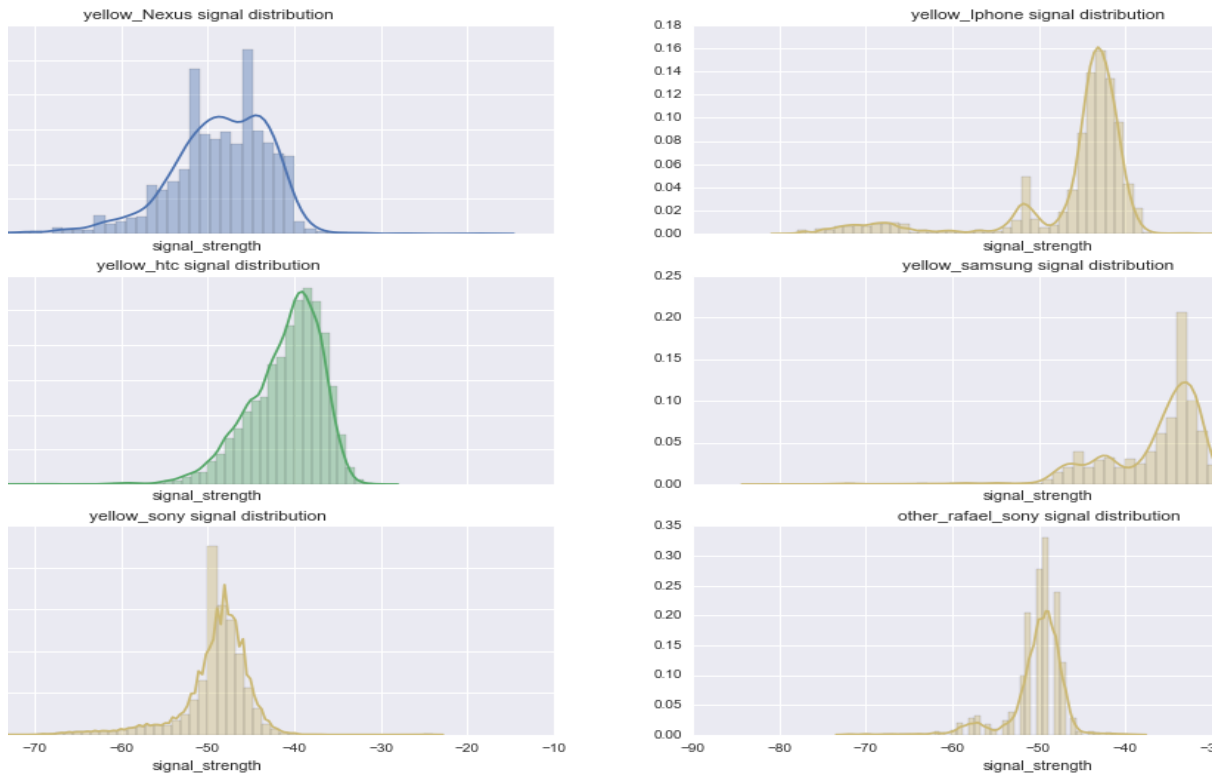
Figure 4.6 - Test case 1 - Phones in the same place

- **Test case #2(Changing location over time)**

Figure 4.7 - Test case 2 - Changing location over time

4.2.3. Q3.What is the effect of multi floor and what the signal ranges are?

- Findings**:**

  - In case of the node in the ceiling of the floor, the results show that the signal strength ranges has a similar values on the floor above (Contorion floor) as Minodes floor (e.g. the Muller node has rceived the same ranges of the singal strenghts values from the phone that is located in Minodes floor as from the same phone when it was located in the floor above(Contorion floor), and a lower values when the phone was placed in the floor down. On the contract, in the case of the node is located in the ground floor (e.g. Fabian node), we got similar range values in the floor down as Minodes floor and different ranges (lower values) in the floor above.

  - To solve the multi-floor probelm, placing the nodes on the ground could be an effective solution.

- Insights and potential use:

  - There is a device offset and could be useful to take it in consideration in the data preprocessing phase especially for the model based approach.

  - The similarity of vendor`s phones behaviours could be helpful in categorizing the phones by vendor regardless the model numbers.

- Test cases:

Distribute the phones at five different locations in Minodes floor for a specific period of time and repeat the experiment on the up floor and the down floor with the approximate locations and for the same period of time.

  - Collect the measurements from a sensor that is located in the ceiling (e.g. Muller node) of Minodes floor.

  - Collect the measurements from a sensor that is located on the desks (e.g. Fabian node) of Minodes floor.

- Test weaknesses:

  - the phones in the down floor were put in locations not exactly as the places in the other floors because the layout in the down floor was completely different and it was not possible to do accurate approximations of the locations.

  - the phones were put in the same places but without order. this means that IPhone for example was placed in a location in Minodes floor that

was different than the other two floors. Therefore, it could be a device offset in this experiment.

- The test was applied only for 5 different phones.

- The test was made only one time because we didn't want to disturb the neighbours.

- Test stregnths:

  - The test measurements were collected by two nodes to show the distance effect (one in the ceiling, another in the roof).

  - The test was applied in real world environment.

- **Test Case#1 (A node on the ceiling)**
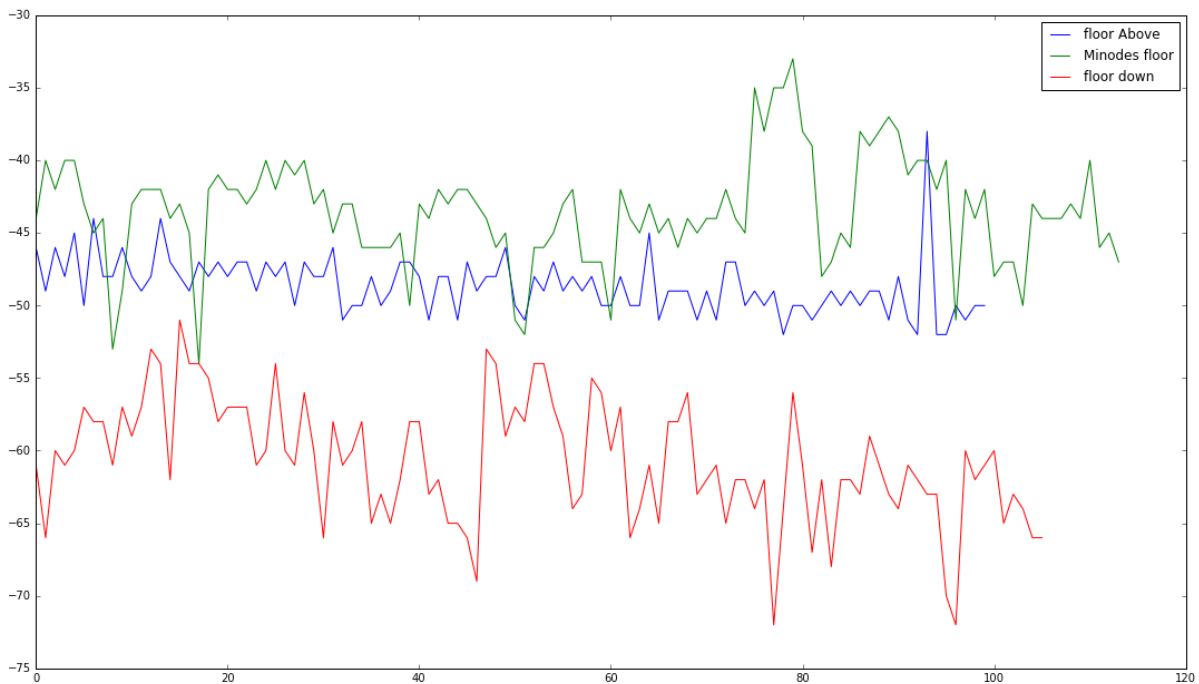
Figure 4.8 - Test case 1 - A node on the ceiling

- **Test Case#2(A node on the floor)**
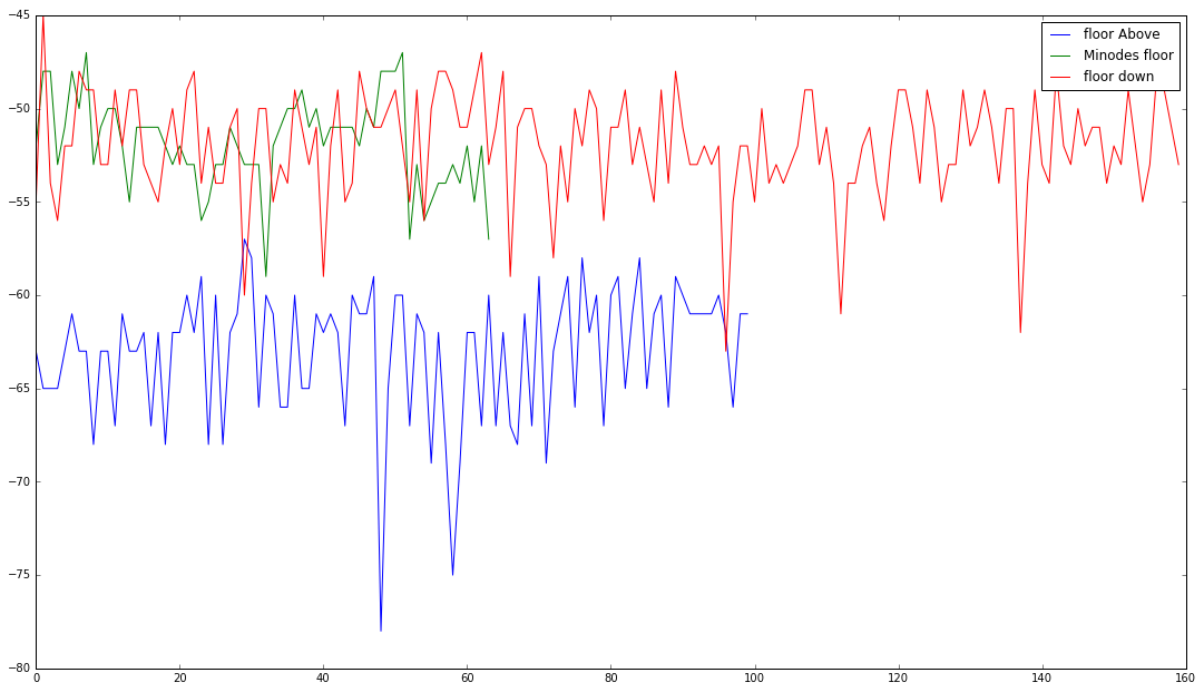
Figure 4.9 - Test case 2 - A node on the floor

4.2.4. Q4. What is the effect of ping delays and is it different by vendors?

Every Wi-Fi-enabled mobile device sends pings (probe requests) to the available Wi-Fi access points (nodes). The client device keeps sending these pings frequently. This frequency of sending the pings could be constant or variant over time based on many factors. One of these factors, for instance, if the phone in active mode or standby mode.

- **Test Case#1 (controlled experiment)**: the measurement phones were used in this experiment inside Minodes building in real world environment to test for the ping delays (frequencies).

- Findings:

  - It seems that the ping frequency varies between different phones vendors and almost the same within the same phone vendor.

  - No clear pattern was found and probably due to the fact that there are many confounding variables that affect the ping frequency.

  - These factors are such as:

  - Standby mode: if the phone is in active mode or standby mode.

- **Battery level:** when the battery is low, the ping frequency became lower in order to reduce the power consumption.

- **WI-FI Connectivity status:** if the phone is connected already to one of the available WI-FI networks.

- The phone vendor which is also affects the ping frequency.

- Insight and potential use:

  - It could be useful to use this insight for both data pre-processing and post-processing purposes.

- Test cases:

  - Putting the measurement phones near specific node and collect node readings.

- Test weaknesses:

  - The test was not able to cover precisely the different factors mentioned above which affects remarkable the ping frequency and the results was not consistent.

  - The test was applied only one time.

- Test strengths:

  - Applying the test to 6 different phones from different vendors.
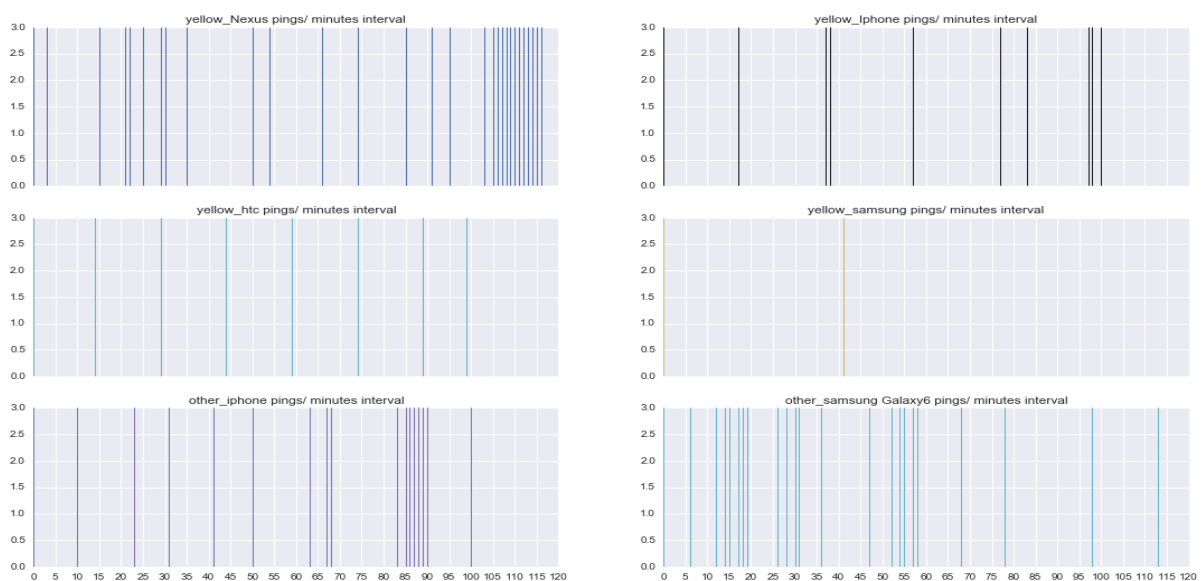
Figure 4.10 - Test case 1 - controlled experiment

- **Test Case #2 (uncontrolled experiment)**

For the purpose of measuring and evaluation the ping delays for different phones. A sample of data was collected from one of the company clients based on a specific period of time and from a restricted randomly chosen node. This sample contains all the Wi-Fi pings from the WI-FI-enabled devices that were in the range of the selected node in the client store. The node that was chosen to collect the data from, is located in a popular area in the store where the clients mostly visit. The main reason for this selection is that the more clients bypass in the area that is covered by the selected node, the more data were collected from more diverse phone models. Hence, a better data analysis for the ping delay case. After filtering out the non-portable WI-FI-enabled devices pings, and after getting the vendor names for all the client devices in the selected data. A lambda function has been defined to calculate the time difference between two consecutive pings by the same client device. This time difference was called delta. To get the overall analysis of the phone vendors, this delta values was grouped by phone vendor.

- Findings:

  - Signals from all WI-FI-enabled devices in the store were collected and that includes fixed devices in the store such as: desktop pcs/tablets/printers, etc. this shows the importance of the pre-processing phase to filter these kinds of data.

  - Apple and Samsung devices show the lowest variation in ping frequency and the highest values of the ping frequency which is very good condition for the data quality.

- Insight and potential use:

  - It could be useful to use this insight for data pre-processing purposes and post-processing purposes.

- Test cases:

  - Collecting the data from a specific store for specific period of time (around one hour) then grouping the data by phone vendor.

- Test weaknesses:

  - No pre-known conditions of all the phones that were in the experiment which provides no insights about the confounding variables that affect the ping delay.

  - The test was applied only for specific period of time which probably does not represent the whole population in a good way.

- Test strengths:

  - The test was applied in a real world environment.

  - The test used a large sample of data.

The test did not control any of the conditions of the phones to show the randomness and to represent the real world environment. (the same point that was considered before as a test weakness).
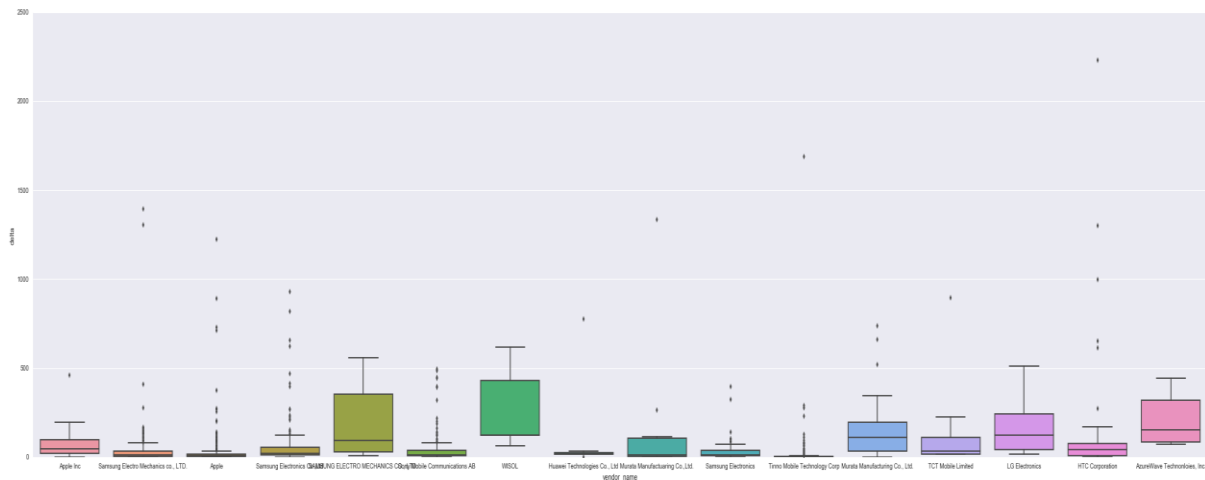
Figure 4.11 - Test case 2 - Uncontrolled experiment

4.2.5. Q5. Is there any effect of the choice of the aggregation method, based on an interval time, on the data quality?

- Findings:

  - Well, there is a remarkable effect on the signal representation by choosing a different methods of the aggregation by interval time. Choosing the "Min" function is the worst option. On the other hand, choosing Max seems to be the most reasonable method since it reduces the interference between signals ranges or it makes the ranges as far as possible between two client devices.

- Insights and potential use:

  - Changing the aggregation method in the Minodes data pre-processing phase to be "Max".

- Test cases:

- Putting two Sony phones within a 2-meter distance relatively to a specific node for one hour and change their locations to make them a bit closer (1-meter) to see the effect of the aggregation method on the signals interference.

- Collect the measurements from the "Muller node" for these two phones.

- Collect the measurements from the "Operation node" for these two phones.

■ Test weaknesses:

- The test was applied only for two different phones.

■ Test strengths:

- The test measurements were collected by two nodes to check the pattern.

- The test was applied in real world environment.

- Two same-vendor phones were used in this test to reduce the device offset effect.

Figure 4.12 - Aggregation method comparison

4.2.6. Q6. What kind of Shielding material that affects the WI-FI signal strength and direction?

- Findings:

  - After trying many materials, no remarkable effect of these materials on the Wi-Fi signals and play no remarkable role in practical in shielding it.

  - No practical solution was found for solving the multi-floor problem that was mentioned before and which is the main catalyst for this experiment.

- Insight and potential use:

  - As mentioned before, there is no practical useful insight from this experiment but it helps to redirect the research of solving the multi-floor problem to a different direction.

- **Test cases:**

- **Test Case #1 - Metal effect**

  - sticking one node(AP) on the wall and then covering it by a metal. Collecting the readings from this node for three of the measurement set phones that were placed in the same place for a specific period of time.

  - The following line chart shows the three received strengths values for these three phones over time.

  - Note: the following line chart shows a dramatic increase in the beginning and that's only because the location of two of the phones was changed during the test.

- Test weaknesses:

  - The test was applied only with the metal material that showed in the following figure without trying different metal effects.

  - The test has been applied on one sensor without trying with another one just to increase the confidence of the results.

  - The metal does not cover tightly the sensor from all directions.

- Test strengths:

  - Appling the test in real world environment.

- Results:

  - No remarkable change in the signal strengths values.

Figure 4.13 - Metal effect 1
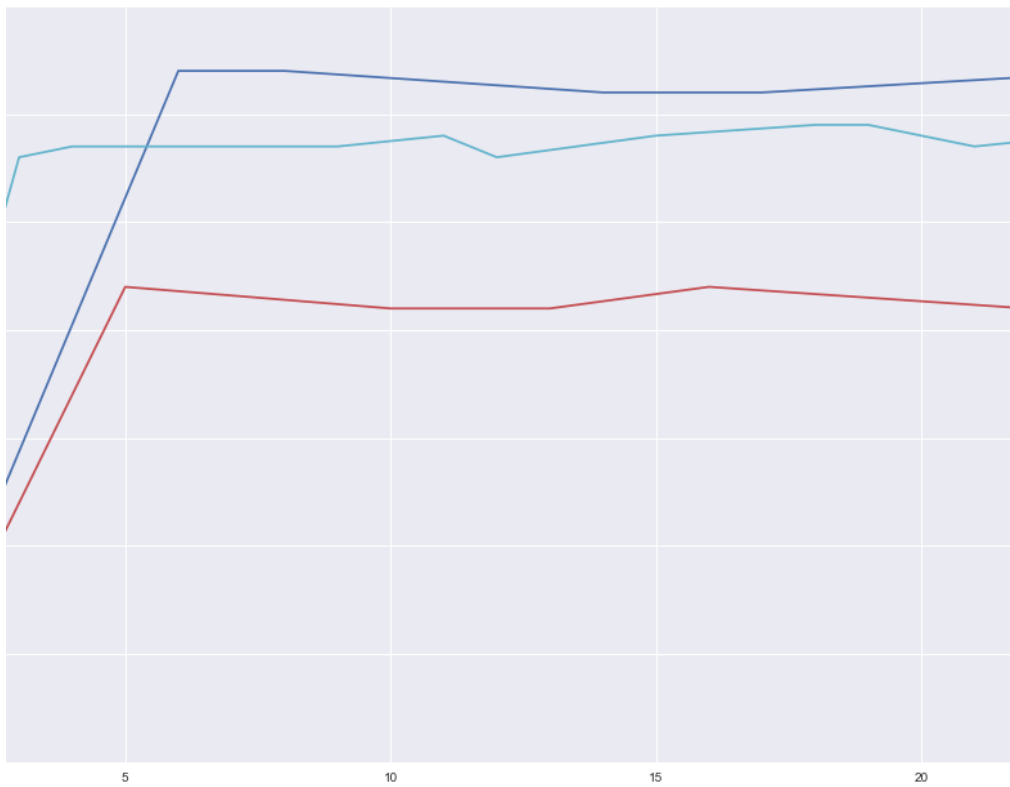
Figure 4.14 - Metal effect 2

- **Test Case #2 - Water effect**

    - sticking one node(AP) on the table and then covering it by the water. Collecting the readings from this node for five of the measurement set phones that were placed in the same place for a specific period of time.

    - The following line chart shows the five received strengths values for these five phones over time.

    - Note: the following line chart shows no values for some phones in some points of the measured time and that is due to the ping delay that was discussed before.

- Test weaknesses:

    - The test has been applied on one sensor without trying with another one just to increase the confidence of the results.

    - The water does not cover tightly the sensor from all directions.

- Test strengths:

    - Appling the test in real world environment.

- Results:

    - No remarkable change in the signal strengths values before and after the water.

Figure 4.15 - Water effect 1
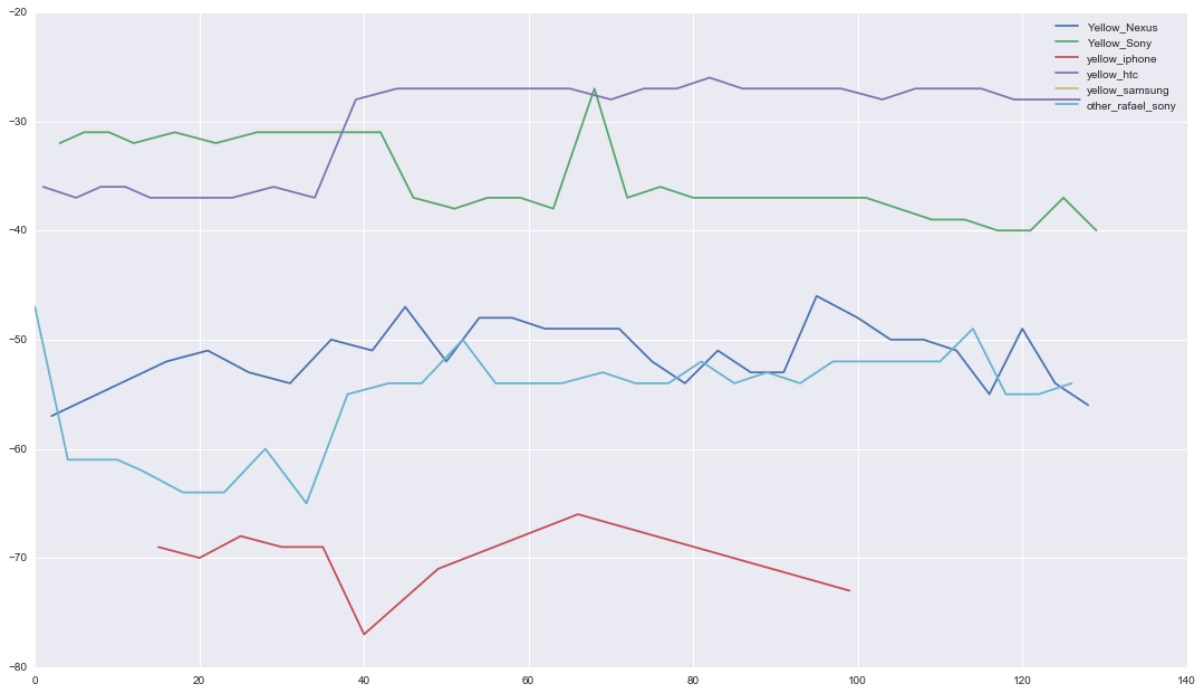
Figure 4.16 - Water effect 2

- **Test Case #3 - Glass effect**

  - Sticking one node(AP) on the table and then covering it by the glass as in the following figure. Then, Collecting the readings from this node for three of the measurement set phones that were placed in the same place for a specific period of time.

  - The following line chart shows the three received strengths values for these three phones over time.

  - Note: the following line chart shows dramatic fluctuation for Sony phone (The personal phone of the researcher) in some points of the measured time and that is due to the changing of the location over time because of making calls during the experiment. Thus excluding the measurement of this phone could be better for analyzing the results. And it was chosen to continue this experiment and not repeating it just to show the real world environment effect.

- Test weaknesses:

  - The test has been applied on one sensor without trying with another one just to increase the confidence of the results.

  - The glass does not cover tightly the sensor from all directions.

- Test strengths:

  - Appling the test in real world environment.

- Results:

  - No remarkable change in the signal strengths values before and after the glass.
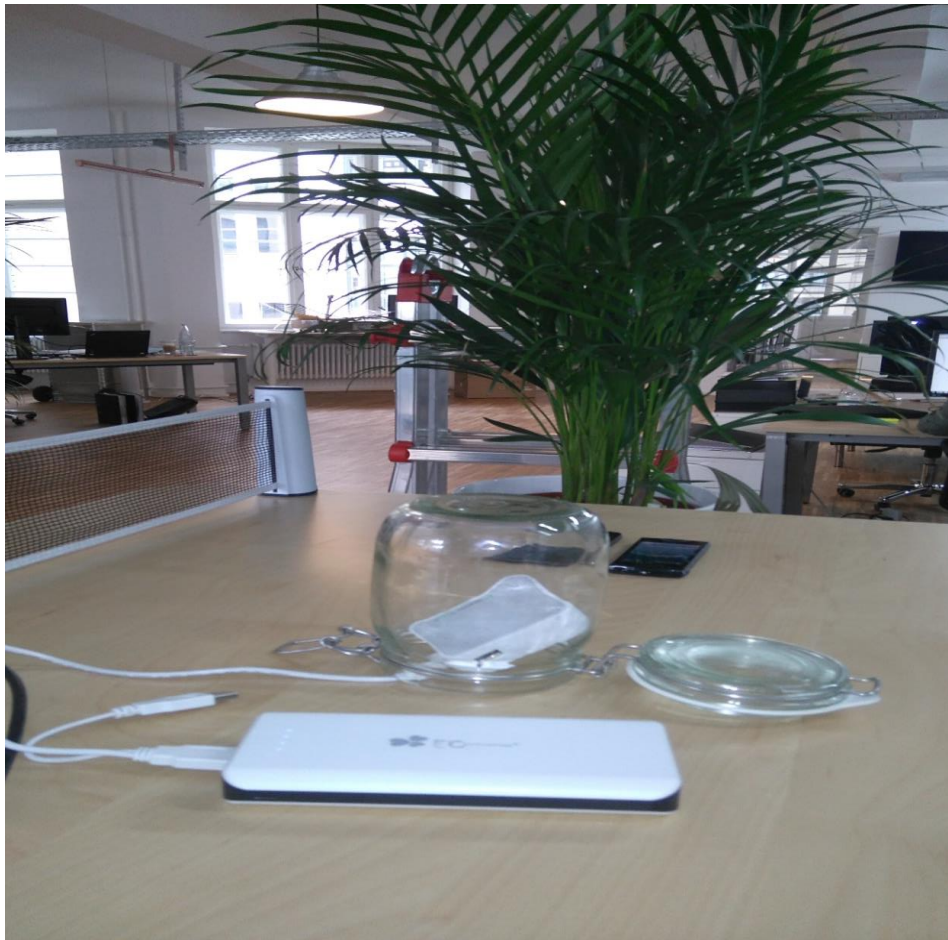
Figure 4.17 - Glass effect 1
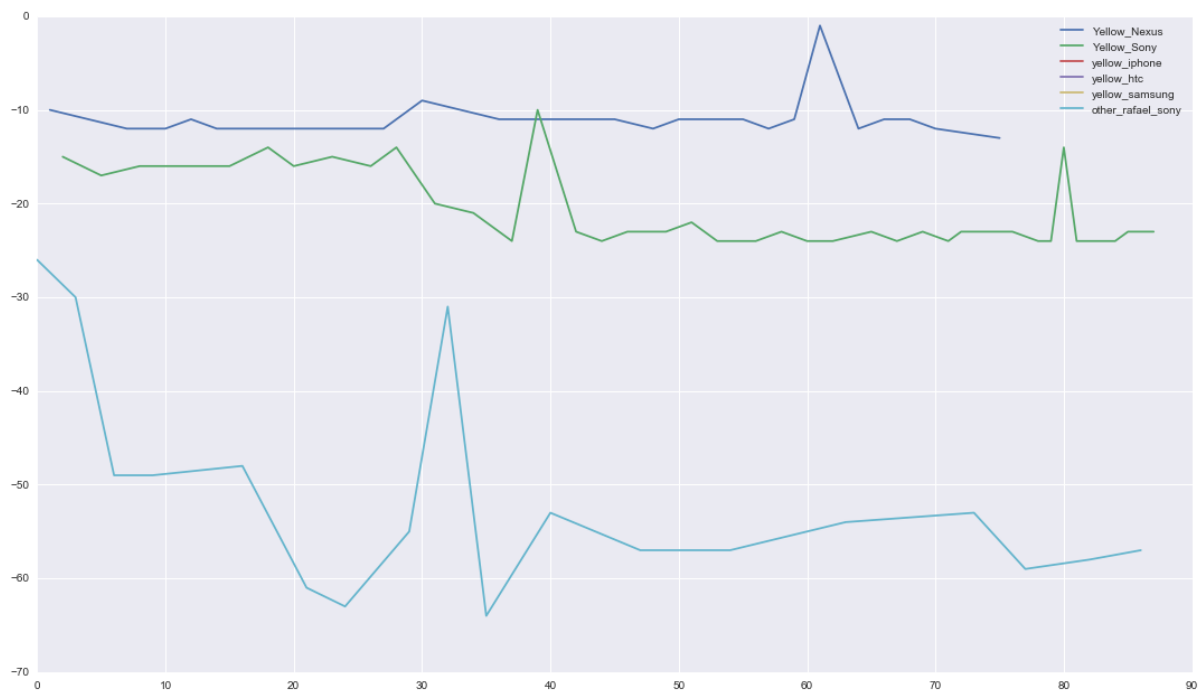
Figure 4.18 - Glass effect 2

- **Test Case #4 - Microwave effect**

  - Given that Microwave oven uses a radio waves and the waves are highly isolated inside the oven, the purpose of this test is to figure out if this isolation holds for WI-FI signals and if so, is it because the device special materials?

  - Putting one sensor inside the microwave oven with closed door. And collecting the readings from the node for two specific phones that were placed in the same location for a specific period of time. The readings were collected before and after putting the node inside the microwave oven to make the comparisons.

  - The following line chart shows the two received strengths values for these two phones over time.

- Test weaknesses:

  - The test has been applied on one time without trying with another sensor on another time just to increase the confidence of the results.

  - No high clear evidence of the true effect.

- Test strengths:

  - Appling the test in real world environment.

- Results:

  - There is huge effect of the microwave oven on the measured signal strength values, the following line chart shows a dramatic decrease in the measured values after putting the node inside.

Figure 4.19 - Microwave effect 1
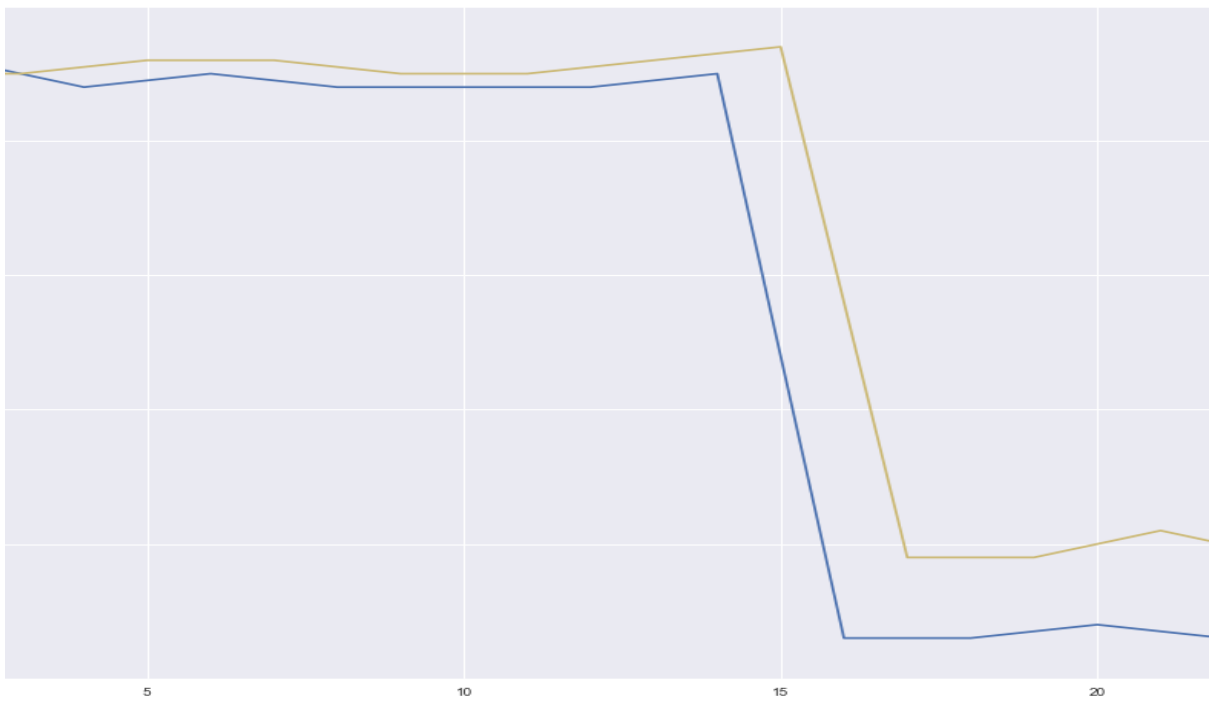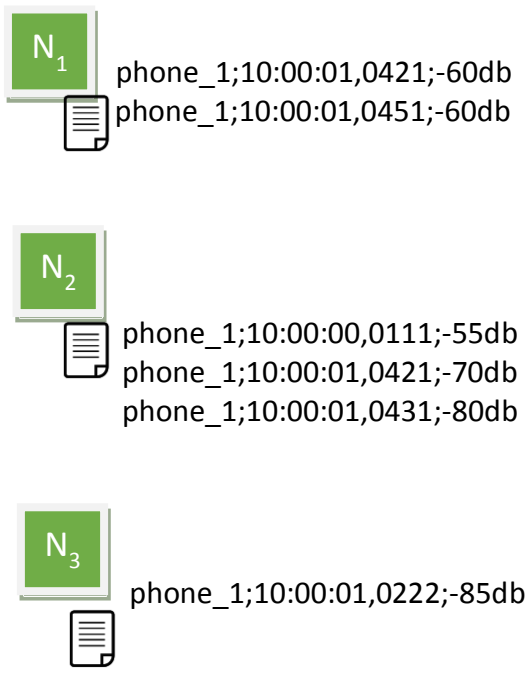
Figure 4.20 - Microwave effect 2

4.3.1. Observation gathering

The first step in the data preprocessing is to collect the data from the files that come from the nodes(APs) and store the data into a structured database as records labeled with the node id as described in the following table:

Table 4-3 - Observations gathering

Observation gathering		
Input	Output	Data Attributes
Files from Observation receiver	N₁ phone_1;10:00:01,0421;-60db phone_1;10:00:01,0451;-60db N₂ phone_1;10:00:00,0111;-55db phone_1;10:00:01,0421;-70db phone_1;10:00:01,0431;-80db N₃ phone_1;10:00:01,0222;-85db	Node_id Mac_ Address Timestamp RSSI value

This step could be improved through using machine learning techniques to detect the best time to transfer the data from the slaves' nodes to the master node to reduce the data loss due to the half-duplex mode.

4.3.2. Aggregating the consecutive phase

The phones usually send a very consecutive pings in a mille-second fraction, these pings usually have different signal strengths (see the following figure). The company current

system (data pipeline) aggregates these consecutive pings using median value that stores in one record that replaces those consecutive records (pings). But after analyzing the signal strength of these consecutive pings, it turns out that getting the ping with the maximum RSS value and ignores the other pings give a better discrimination ability for the signal strength and lead to better positioning prediction using different approaches. The following table shows an example of the input and the output of this step.
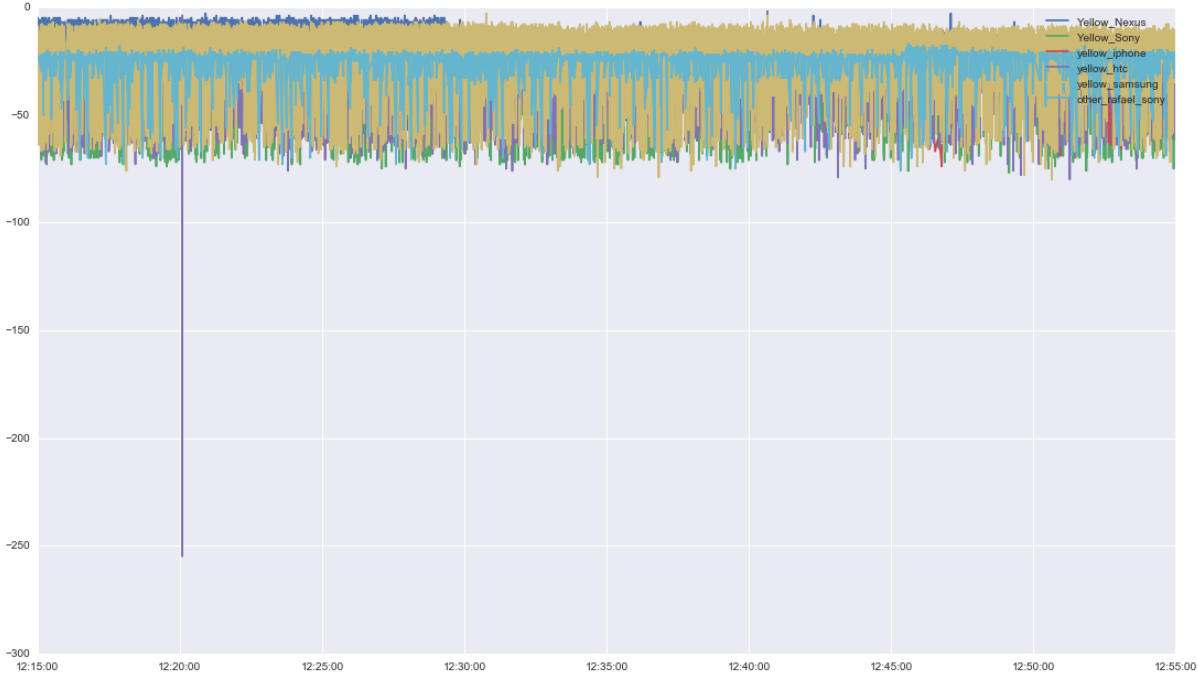
Figure 4.21 - Raw data

Table 4-4 - Signals aggregating

Signals aggregating (bursting)		
Input	Output	Data Attributes
N₁ phone_1;10:00:01,0421;-60db phone_1;10:00:01,0451;-60db N₂ phone_1;10:00:00,0111;-55db phone_1;10:00:01,0421;-70db phone_1;10:00:01,0431;-80db N₃ phone_1;10:00:01,0222;-85db	N₁ phone_1;10:00:01,0421;-60db N₂ phone_1;10:00:00,0111;-55db phone_1;10:00:01,0421;-70db (observation with strongest signal within burst) N₃ phone_1;10:00:01,0222;-85db	Node_id Mac_ Address Timestamp RSSI value
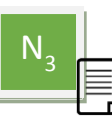
This step could be improved via the following:

- Doing the bursting phase on a mille-second fraction so we can collect more records during the training phase to improve the accuracy.
- Choosing the appropriate aggregate function (Maximum, Minimum and Median or a custom one).

4.3.3. Fingerprinting

This step is necessary to collect all the aggregated signal strength values (RSS) from all the nodes in a specific store in a specific timestamp for a specific phone. The data will be in matrix-like form where the columns represents the installed nodes in a specific store. And each row contains RSS values for a specific client device in a specific timestamp (It is worth mentioning that the hashed mac address represents the specific device ID). This row is called a fingerprint. Thus, the whole data set will be an array of fingerprints.

Table 4-5 - Fingerprinting

Fingerprinting		
Input	Output	Data Attributes
N₁ phone_1;10:00:01,0421;-60db N₂ phone_1;10:00:00,0111;-55db phone_1;10:00:01,0421;-70db N₃ phone_1;10:00:01,0222;-85db	*Per (phone_1, timestamp)* $FP1 \begin{pmatrix} -60 \\ -70 \\ -85 \end{pmatrix}$ $FP2 \begin{pmatrix} NA \\ -55 \\ NA \end{pmatrix}$	Mac_ Address Timestamp RSSI Vector

4.3.4. Imputing the missing values

The fingerprint in the data set may contains missing values for some nodes. This could happen because of:

- Node failure due to technical issues.

- Network connection problem where the node was not able to send its data to the server.

- Out of range problem where the client location is far away from some nodes that could not capture any signal from the client device.

Table 4-6 - Missing values imputing

Incomplete fingerprint imputing		
Input	Output	Data Attributes
Per (phone_1,timestamp)	*Per (phone_1,timestamp)*	Mac_ Address
		Timestamp
FP1 $\begin{pmatrix} -60 \\ -70 \\ -85 \end{pmatrix}$	*FP1* $\begin{pmatrix} -60 \\ -70 \\ -85 \end{pmatrix}$	RSSI vector
FP2 $\begin{pmatrix} NA \\ -55 \\ NA \end{pmatrix}$	*FP2* $\begin{pmatrix} NA \\ -55 \\ NA \end{pmatrix}$ (crossed out) (or filling using imputing strategy)	

This step could be improved through the following:

- Missing values imputing using a classifier.

- Missing value imputing using mean/median values.

- Ignoring/removing the fingerprint when number of missing values reach a predefined threshold.

4.3.5. **Scaling**

For better performance of any modeling technique, a scaling of the features is such an important step. Many scaling methods could be used such as: Z-scoring and range scoring.

Table 4-7 - Scaling

Scaling		
Input	Output	Data Attributes
FP1 $\begin{pmatrix} -20 \\ -10 \\ -10 \end{pmatrix}$	*FP1* $\begin{pmatrix} 0.8 \\ 0.4 \\ 0.4 \end{pmatrix}$	Mac_ Address Timestamp RSSI vector

This step could be improved through the following:

- De-logarithmic scaling of the signal strength: RSS value(in dBM) is logarithmic value of base 10, so de-logarithmic the value could be an effective solution.
- Apply range scaling instead of z-scoring.
- Using machine learning technique to choose the appropriate scaling function or even without scaling.

4.3.6. Getting the vendor names of the client devices

To provide analysis on the data by the client device vendor, the phone vendor code,which is encoded in the client device mac address, was used. The phone vendor OUI or organisationally unique identifier is a 24-bit number or basically the first three octets of a client device MAC address. It identifies a vendor or manufacturer of a client device. An example of a phone vendor:

OUI   Vendor {Company}
84:38:38   SAMSUNG ELECTRO-MECHANICS(THAILAND)

To look up the phone vendor names of all the phones MAC addresses in our sample data, an open public API has been used[1]. It provides the commercial name of the phone vendor when the MAC address or the phone vendor code (OUI) is entered.

---

[1] Used API URL: http://www.macvendorlookup.com/api

4.3.7. Non-portable WI-FI-enabled devices filtering

The collected data from any node in any store may contains a good deal of records (pings) from a fixed non portable WI-FI-enabled devices. These devices include, but not limited to, PCs, tablets, printers, scanners that have a WI-FI network interface card is installed and configured to send probe requests to connect to the available Wi-Fi networks in the coverage area. These devices could be located inside the store owned by the shops dealers or outside the store from the neighbouring buildings. To filter out the pings that came from these devices. The records (pings) that have a constant or near constant signal strength(RSS) value over the predefined monitored time are considered as pings generated from fixed (non portable) client devices. Filtering out this data will reduce the sample data for the analysis and could reduce the performance of the modelling. But, but businesswise, this filtering is very important step and should be done because it helps giving a better estimation of the number of store visitors and visiting routes and duration. And that is obviously the main purpose and the core of the retail analytics and that is why these WI-FI positioning approaches are applied in this field.

The following figure shows a sample of the collected data from a specific store. It shows the variation of signals grouped by device vendors including the devices like tablets, pcs,printers, etc.
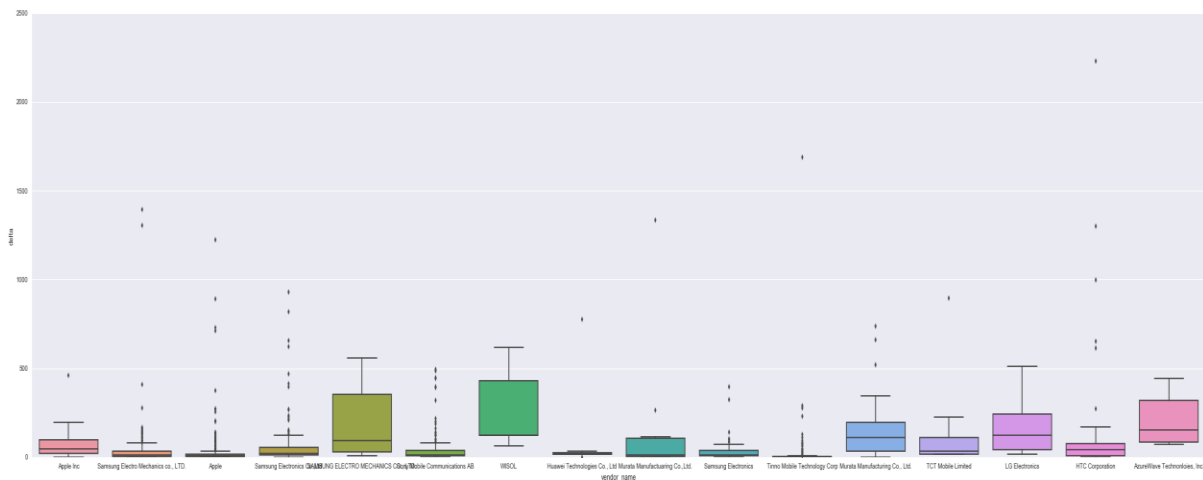
Figure 4.22 - Unfiltered grouped by vendor data

4.3.8. Zone re-numbering

As it is explained before, during the setup phase, the zones were giving numbers arbitrarily without specific order. The numbering of the zones is important step, only, for the pattern recognition based approach because the zone number (label) is the target class and the ordering between classes make a big difference. A new non-arbitrary approach has been proposed which takes in consideration the physical order of the zones according to their locations in the floor plan.

4.4. Feature engineering

As a part of feature engineering, new features have been introduced as predictors for the model.

- N_active_nodes: this represents the number of active nodes in each fingerprint. The active nodes are the nodes that have received signal for a specific probe request and transferred the RSS values into the central server as a raw data to form later the fingerprint data.
- N_strong_nodes: this represents the number of strong nodes in each fingerprint. The strong nodes are the nodes that have received strong signal above a certain preconfigured threshold (e.g. -35) for a specific fingerprint.
- Phone vendor: as shown before in the exploratory data analysis, the phone vendor affects remarkably the signals strengths and their variations. Adding this feature will help improve the accuracy a great deal.
- Zone number: as mentioned before, the zone numbering approach had been proposed to change the previous arbitrary way of numbering. This approach reflects the actual place order of the zones in the store. And that converts the target class from unordered labels to ordered labels where the order between the zones matter.

4.5. Modeling-based mapping Approach implementation

As previously mentioned, this approach is popular for indoor positioning systems. And the company are using this approach as the only method and the core of its product.

The training set is collected during the setup phase (as explained before) by the operation team. This training set has known labels which are the zone numbers (target class). It contains all the signals (probe requests) that have been sent from the measurement set phones, that have been used by the operation team during the calibration phase, from every place in the store during the whole time period of the calibration phase. It is worth mentioning again that after the pre-processing of the collected data, a matrix form of RSS values from all the installed nodes has been formed (see the table below). Each record of this matrix represents one probe request (called a fingerprint) with all RSS values from all the installed nodes.

Probe request ID	RSSI value node 1	RSSI value node 2	RSSI value node 3	RSSI value node 4	-----	RSSI value node 150	Zone ID (target class)
1	-70	-50	-45	-30	----	-88	15
2	-50	-45	-66	-42	----	-30	4

3	-53	-42	-67	-39	---	-37	4
---	---	---	---	---	----	---	---

Those fingerprints are considered as predictors (independent variables). It is clearly seen that the number of these predictors depends on the number of the nodes that are installed in the location. That is because each one of the variables represents the RSS value that has been received by a specific node in the specified timestamp. These independent variables (features) are named by their nodes IDs (See the figure below).

```
# feature names in this data set refering to node ids
features = ['5', '74', '116', '141', '32', '137', '127', '71', '38', '128', '4', '59', '77', '55', '80', '35', '78', '66', '112',
'9', '149', '86', '146', '27', '24', '145', '95', '111', '139', '132', '73', '15', '106', '23', '40', '1', '136', '97', '99',
'81', '117', '62', '46', '144', '108', '91', '122', '39', '64', '12', '94', '113', '63', '25', '14', '34', '44', '47', '51',
'48', '49', '22', '18', '21', '19', '93', '138', '114', '37', '120', '58', '85', '104', '41', '56', '90', '30', '107', '92',
'31', '76', '131', '142', '57', '143', '87', '135', '100', '50', '115', '36', '124', '70', '98', '110', '67', '72', '88', '103',
'96', '11', '52', '65', '13', '126', '118', '2', '68', '7', '133', '102', '101', '33', '109', '26', '69', '8', '28', '123',
'121', '79', '53', '75', '20', '130', '43', '89', '3', '17', '45', '105', '16', '125', '29', '84', '150', '54', '42', '140',
'60', '82', '61', '119', '129', '10']
```

Figure 4.23 - Feature names, [Minodes]

The company has been using a machine learning model (random forest) to predict the unknown zone numbers for the new visitors. The accuracy of this model is considered as a reasonably good. What It is meant by the accuracy is the goodness of the model to predict the visitor location inside the store and that means finding the right zone that the customer is inside in a specific time.

In this study it had been tried to use different machine learning techniques to estimate the customer location.

- Gradient boosting.
- Lazy learning methods has been used and specifically k-nearest neighbor algorithm.
- Support vector machines.
- Naïve based multinomial model.
- Random forest.

The company location was used as an experimental field to check the performance of these different machine learning techniques. Several attempts for tuning the parameters for these machine learning algorithms to reach a good model and to avoid overfitting were done.

4.6. PROXIMITY DETECTION MAPPING APPROACH IMPLEMENTATION

This method is also called Cell-of-Origin. A new cell of origin mapper had been introduced, implemented and tested. As was mentioned before, it is quite important to install one node

in each zone. Also it is recommended to setup the node in the center of the zone. These recommendations have been taken in consideration by the operation team for the new installation for clients.

The basic principle in this method is to determine the location of the visitor by the location of the node that recorded the strongest signal from the client device. It is clearly how simple it is, because there is no need for any training set. Thus, no need for any historical data nor remembering the previous states. The cell of origin mapper can act and provide results in a real-time basis.

The figure below shows some implementation of the cell of origin mapper. The function showed in the picture takes as an input a specific phone (phone MAC address) and the store id and the nodes ids within this store. This function calculates, on a real-time basis and again without any historical data and without any learning system behind it, the total visitor journey inside the store and within the specified period of time. It reports the zones numbers that the visitors visit with the timestamp for this visit.

```python
def cell_of_origin_individual_total_visit(store_key,node_ids, phone_hash,
                                          phone_name, start_time,end_time_delta_mins,
                                          interval_time_mins):
    periods=(end_time_delta_mins) /interval_time_mins
    #(end_time-start_time) /np.timedelta64(1, 's')
    visits=[]
    end_time=start_time+ timedelta(minutes=interval_time_mins)
    for period in range(0,int(periods)):
            location=get_node_location(cell_of_origin_individual_mapper(store_key,node_ids,
                                          phone_hash, start_time, end_time))
            #print(str(start_time)+"\t"+str(end_time)+" " +location)
            visits.append([str(start_time),str(end_time),location])
            start_time=end_time
            end_time=start_time+ timedelta(minutes=interval_time_mins)
    print(pd.DataFrame(visits))
    pd.DataFrame(visits).to_csv("output_total_visit_cell_of_origin "
                                +phone_name+" "+str(start_time.date())+".txt")
```

Figure 4.24 - Cell of origin - few implementation

The following figure shows an example of the results that were produced by the mapper in a real world experiment. The company building was used for this experiment and the mobile device of one of the company employees was tracked. This example shows the mapper usage as a tracking system for one client device over time just to show the capability of this approach and how simple it is to implement.

3,2016-03-15 09:45:00,2016-03-15 10:10:00,   Kitchen

4,2016-03-15 10:10:00,2016-03-15 10:35:00,   Eva office

5,2016-03-15 10:35:00,2016-03-15 11:00:00,   Meeting near library

6,2016-03-15 11:00:00,2016-03-15 11:25:00,   Meeting near library

7,2016-03-15 11:25:00,2016-03-15 11:50:00,   Meeting near library

8,2016-03-15 11:50:00,2016-03-15 12:15:00,   Alex kouth

9,2016-03-15 12:15:00,2016-03-15 12:40:00,   Kitchen

10,2016-03-15 12:40:00,2016-03-15 13:05:00,   Kitchen

11,2016-03-15 13:05:00,2016-03-15 13:30:00,   Kitchen

12,2016-03-15 13:30:00,2016-03-15 13:55:00,   Kitchen

13,2016-03-15 13:55:00,2016-03-15 14:20:00,   Muller place

14,2016-03-15 14:20:00,2016-03-15 14:45:00,   Kitchen

15,2016-03-15 14:45:00,2016-03-15 15:10:00,   Muller place

16,2016-03-15 15:10:00,2016-03-15 15:35:00,   Library

17,2016-03-15 15:35:00,2016-03-15 16:00:00,   Muller place

Figure 4.25 - Cell of origin mapping example

5. RESULTS AND DISCUSSION

Firstly, an important outcome of this study is that it helped in better understanding the behavior of the WI-FI signals which is the backbone of the positioning approaches. It helped in understanding the different variations that affect the WI-FI signals such as reflections, multi paths, etc. Also, that benefits in understanding the underlying process of data generation which leaded to better data preprocessing. It was also clear that this study results significantly in deep analysis of the relationships between different factors that affect the positioning system. And that made a lot of difference especially in introducing new features that were not taken in consideration before.

Secondly, during this study, many attempts were done mainly, as explained before, to improve the currently used fingerprinting based positioning approach. And that included building many machine learning models that were trained after applying the preprocessing of data and after applying the feature engineering (see the following table).

Table 5-1 - Fingerprinting based approach - machine learning modeling results

Model	Accuracy	Notes	Time
KNN	0.4460920606	15 neighbours uniform without scaling	44 minutes
KNN	0.396184739	15 neighbours distance without scaling	50 minutes
KNN	0.4546339203	15 neighbours distance with scaling	45 minutes
KNN	0.4084182885	15 neighbours uniform -200 the very week signal , adding 200 to all the features values	45 minutes
xgboost	[41] train-merror:0.392314 val-merror:0.505375	test size 30% param['objective'] = 'multi:softmax' # scale weight of positive examples param['eta'] = 0.1 param['max_depth'] = 6 param['silent'] = 1 param['nthread'] = 4 param['num_class'] = 139 param['eval_metric']='merror' param['max_delta_step'] = 1 watchlist = [(xg_train,'train'), (xg_val, 'val')] num_round = 600	
xgboost	[130] train-mlogloss:0.793833 val-mlogloss:1.343777	softprob, mlogloss, 600 rounds, 70/30 train test split	one hour and half
SVC	0.4981927711	one versus one decision_function_shape='ovo'	two hours
SVC	0.499382144	one versus rest ovr	94 minuttes
SVC	0.4533209762	linearSVC c=1.0	140 minutes
Naive Bayes multinomial	0.35	multinomial alpha=1	

The results show that the gradient boosting outperformed significantly the company`s currently used machine learning system. But it is also important to mention that a large part of the accuracy improvement was due to the additional steps that were taken in the data preprocessing and feature engineering phases.

Finally, a significant output of this study is that it introduced a new simple online mapper based on cell of origin technique which was used together in mix with the fingerprinting based technique to provide better accuracy of the positioning system overall. This mix can be seen as a hybrid approach between two methods. Also, as another benefit from the new mapping technique, the cell of origin mapper was used also as a post-processing validation technique of the other approach to validate the positioning outcomes and to correct them when it is needed.

6. CONCLUSIONS

Using WI-FI technology for positioning and providing location based services has become popular in the recent years. Even though that was not the purpose of this technology when it was designed. Different approaches for positioning based on WI-FI show a good performance in terms of the accuracy of the localization estimation. Each approach has advantages and disadvantages and no one approach can outperform the others in all cases.

Not surprisingly, some factors in some cases play an important role more than other factors in choosing the best approach. Obviously, the cost-efficiency of the solution is one of the most important factors when it comes to choosing practical solutions to the market.

In the retail analytics case study that is presented in this document, practical results from the experiments that were held, show a very reasonable performance of the fingerprinting based approach. And it shows also from the cost efficiency perspective, which is very important practically, that it is a very effective method because, as discussed before, no high-cost hardware resources are needed for the implementation. Which is certainly a big problem especially in the large scale where large number of nodes are needed to cover sufficiently the customer`s store/shop. And that is why, no surprisingly, the company has used this approach as the core positioning system for their retail analytic product.

Finally, what this study shows and suggests, after practical real world experiments, is that even mixing one approach (the fingerprinting) with another cost effective approach (Cell of Origin) could lead to better results with no additional cost. Of course, it is worth mentioning here the work that has been done during this study on the preprocessing and feature engineering. Which also increase the accuracy of the positioning a great deal.

7. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

It is worth mentioning that during this study, other approaches for the indoor positioning were tested in real world experiments. But there were some limitations and obstacles.

7.1. ANGLE OF ARRIVAL (AOA) APPROACH LIMITATION AND FUTURE WORK

During this study, an attempt to do some experiments with this approach was started. directional antennas (See the figure below) were used in the beginning to study the effect of the antenna`s direction on the received signal strength. After connecting these antennas to the nodes as an external antenna to act as a receiver, no remarkable effect was recorded on the received signal strength after changing the antenna direction (the receiver). The failure of this experiment could be likely due to the hardware (antenna) quality (Halperin, Hu, Sheth, & Wetherall, 2010) that have been used in this test.

Figure 7.1 - Directional antenna

Also, a MIMO- Enabled WI-FI network card had been purchased, specifically Intel Wi-Fi Wireless Link 5300 802.11n MIMO, and prepared to be used with the Linux based open source 802.11 measurement and experimentation platform [2]. But for technical reasons the machines (at least two machines needed to install the MIMO – enabled WI-FI cards) were not ready and equipped with Linux-based framework during the study planned time.

As a future work, making sure continuing working on these two experiments because there is a good potential of this approach for indoor positioning (Xiong & Jamieson, 2013) (Rong & Sichitiu, 2006) as was mentioned before in the theoretical study.

Figure 7.2 - MIMO-enabled WI-FI card [Amazon]

7.2. TIME DIFFERENCE OF ARRIVAL LIMITATIONS AND FUTURE WORK

During the study, experiments were planned to test the time difference of arrival approach as another approach for the indoor positioning. The highly accurate measurement of the timestamp was needed for using this approach because of the high speed of the WI-FI signals (near speed of light). At first it was tried to use the NTP time that the mesh network use to synchronize all nodes timestamps. But unfortunately the NTP time is not accurate enough (Microseconds accuracy level) to measure the time of flight of the Wi-Fi signal, even though NTP can achieve better than one millisecond accuracy under ideal condition (Martin & Kasch, 2010).

For that reason, a GPS receiver was purchased (See the following figure) and the purpose was to use the highly accurate GPS time (Nanosecond accuracy level) to setup the NTP time server. But unfortunately the limitation was that the purchased GPS device was not supporting the 1PPS (one pulse per second) output which is needed to setup the NTP time server [3]. For this technical reason, the experiment was stopped and planned as a future work to find a suitable GPS receiver in the market who supports this feature and then it will be able to continue experimenting this approach.

---

[2] CSI tool URL : http://dhalperi.github.io/linux-80211n-csitool/
[3] GPSD lbrary config URL: http://www.catb.org/gpsd/gpsd-time-service-howto.html

Figure 7.3 - GPS receiver [Google images]

7.3. MODELING BASED APPROACH LIMITATIONS AND FUTURE WORK

The training set that had been collected during this study to build the machine learning models contains only the fingerprints from the measurement set phones (5 phones). This could be seen as a limitation since the measurement set is a small sample from large set of phones in the market that could be owned by the visitors. Another limitation is the ability to test the model effectively since collecting the test set also require the same effort as collecting the training set and it requires an effort by the operation team. Also, during this study, the training set was collected in ideal environment where the store was empty from visitors (weekends). This could be seen as a limitation as well since the real world environment affects a lot the WI-FI signals due to reflection, line of sight issues.

As a future work, it is important to focus on finding cost effective ways to measure constantly the performance of the predictive model over time. Because as explained before, due to many reasons (e.g. moving furniture) the performance of the model could be affected a great deal. So finding new ways to measure the performance constantly without the need to repeat the costly calibration phase again.

7.4. PRE AND POST PROCESSING LIMITATIONS AND FUTURE WORK

During the study, it was clear the negative effect of the so-called multi-floor problem. This problem appears due to the similarity between signal measurements in two consecutive floors. As a result, it was confusing to tell in which floor the visitor was. And that is because the nature of the WI-FI signals. A lot of experiments were done during this study to find a shielding material (or any other solution) to redirect the Wi-Fi signal to solve this problem but all of them failed. So a lot of work in the future should focus on finding cost-effective shielding materials to prevent or at least reduce the effect of the multi-floor problem.

Additionally, a significant amount of work could be done in the future using post processing techniques as a validation of the accuracy of the mapper. This is not only including the modelling based mapper but all the other approaches as well. And it does not mean only the post processing on a real time basis but also the batch-processing or offline basis. For example, using the dead reckoning technique (Jain, 2014) or Kalman filter (Honkavirta, Perälä, Ali-Löytty, & Piché, 2009) to validate the results from the mapper.

8. REFRENCES

Curran, K., Furey, E., Lunney, T., Santos, J., Woods, D., & McCaughey, A. (2011). An evaluation of indoor location determination technologies. *Journal of Location Based Services*, *5*(2), 61–78.

El-Rabbany, A. (2002). *Introduction to GPS - The Global Positioning System*.

Farid, Z., Nordin, R., & Ismail, M. (2013). Recent Advances in Wireless Indoor Localization Techniques and System, *2013*.

Gast, M. (2002). *802.11 Wireless Networks. The Definitive Guide*.

Halperin, D., Hu, W., Sheth, A., & Wetherall, D. (2010). 802 . 11 with Multiple Antennas for Dummies. *Computer Communication Review*, *40*(1), 19–25.

Headquarters, A. (2008). Wi-Fi Location-Based Services 4.1 Design Guide, (6387).

Henniges, R. (2012). Current approaches of Wifi Positioning. *IEEE Conference Publications*, 1–8.

Honkavirta, V., Perälä, T., Ali-Löytty, S., & Piché, R. (2009). A comparative survey of WLAN location fingerprinting methods. *Proceedings - 6th Workshop on Positioning, Navigation and Communication, WPNC 2009*, 243–251.

Jain, R. (2014). Survey of Wireless Based Indoor Localization Technologies, 1–17.

Martin, J., & Kasch, W. (2010). Network Time Protocol Version 4: Protocol and Algorithms Specification, 1–110.

Navarro, E., Peuker, B., Quan, M., Clark, A. C., & Jipson, J. (2010). Wi-Fi Localization Using RSSI Fingerprinting. *Test*, 1–6.

Ozsoy, K., Bozkurt, A., & Tekin, I. (2013). Indoor positioning based on global positioning system signals. *Microwave and Optical Technology Letters*, *55*(5), 1091–1097.

Rong, P., & Sichitiu, M. (2006). Angle of arrival localization for wireless sensor networks. *… Ad Hoc Communications and Networks, …*, 9.

Xiong, J., & Jamieson, K. (2013). ArrayTrack: A Fine-Grained Indoor Location System. *USENIX Symposium on Networked Systems Design and Implementation*, (279976), 71–84.