Machine Learning based Indoor Positioning

Baldeep Singh

Submitted for the Degree of Master of Science in

Machine Learning

Department of Computer Science
Royal Holloway University of London
Egham, Surrey TW20 0EX, UK

September 2, 2019

Declaration

This report has been prepared on the basis of my own work. Where other published and unpublished source materials have been used, these have been acknowledged.

Word Count: 14868

Student Name: Baldeep Singh

Date of Submission: 02/09/2019

Signature: Baldeep Singh

Abstract

Wi-Fi fingerprinting is one of the indoor localization techniques which uses the intensity of the Received Signal Strength (RSS) from the wireless access points distributed around the building as the measurement for determining the receiver's position. In the past, various machine learning algorithms have been used to resolve the indoor positioning problem which has resulted in different accuracies and processing time. This dissertation presents a localization model employing K-Nearest Neighbours (KNN) regression and K-means clustering to give a much better accuracy of the user location.

After implementing the hybrid model, the model is experimented with a public database that was collected at Jaume I University in Spain. The results show that the hybrid model has outperformed the individual KNN model by 120%, achieving the position accuracy of 5 meters in contrast to the 11 meters accuracy achieved by the solo KNN model deployed with an optimal value of k. While the KNN model improves the performance by 90% when the value of k is optimised from k=300 to k=5, the KNN and k-means hybrid algorithm further enhances the localization accuracy.

The position of the targets were determined by first predicting the cluster it belongs to and then locating the coordinates by applying KNN only to the predicted cluster. This indicates that the error difference increases gradually with an increase in the number of positions that have been entered in the database. Thus segregating them into groups or clusters would result in a decrease in error. In this dissertation we use the Cumulative Distribution Function (CDF) plots for comparing the accuracy of these models.

Contents

1 Introduction

This chapter first explains the indoor positioning problem, the background research that has been done in this field. Next, our research objective behind this project, the assumptions of study, the overview of approaches and the main contributions are presented. Finally, the structure of this dissertation is outlined.

1.1 What is indoor positioning? Why is it needed?

The best way to explain Indoor Positioning Systems (IPS) is that it's like GPS, but for indoor environments.

It is a mechanism for determining the location of an object in an indoor space using lights, radio waves, magnetic fields, wireless or sound signals, or other sensory information. In the literature, this process is usually termed differently as radiolocation [1], position location [2], geolocation [3], location sensing [4], or localization [5]. This dissertation will primarily use localization, but all of these terms are used interchangeably throughout the document.

A system deployed to determine or estimate the location of an entity is called a position location system or positioning system. The term positioning system will be used to represent the system throughout this document. A wireless indoor positioning system refers to a wireless network infrastructure that provides indoor location information to any requesting end user. A set of coordinates or reference points within the predefined space is typically used to indicate the physical location of the entity. For instance, the global positioning system (GPS) uses the latitude, longitude, and altitude as the coordinates of an entity on the Earth's surface. On the other hand, an indoor positioning system may combine a floor number, room number, and other reference objects to represent an entity's position. Note that the term position and location are used interchangeably even though the first term has a smaller scope than the second term.

Although the technology is newer than GPS, IPS is quickly gaining momentum in places like shopping malls, hospitals, airports and other indoor venues where navigation and other location-based services are required. The applications of indoor location information are not limited to tracking the location of users and objects in both emergency and normal situations. Concierge services enable users to become aware of the nearest supporting facilities. For example, in an office automation system a document can be automatically printed to the closest printer near a mobile user. If a person wearing a location device is not present at his desk, an incoming phone call can be forwarded to the nearest telephone set [6]. In the field of robotics, a robot can navigate by itself using the assistance of an indoor positioning system [5]. Smart home applications such as multimedia appliances that forward multimedia stream to the nearest video screen can be achieved with a home positioning system [7]. These examples are just some emerging location related applications.

So why doesn't GPS work indoors? GPS systems loses its effectiveness in an indoor environment because the weak signals from satellites that GPS rely on are easily blocked by a roof or walls. The result is a significant drop in accuracy when

the user enters a building. That is where the indoor positioning system (IPS) comes into the picture.

There are a number of different technologies that can be used for indoor positioning, some of which we'll cover in detail in this dissertation are:

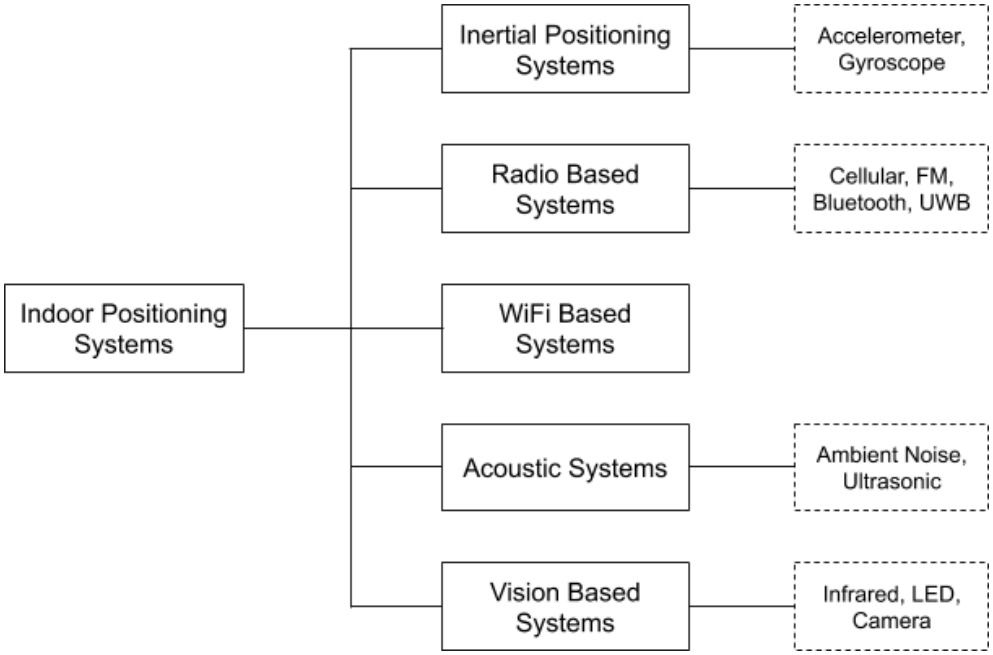


Fig. 1.1: Types of Indoor Positioning Systems

- **Inertial positioning systems**
  - These are the accelerometer or gyroscope based systems that can detect the general location of a person or object at room level within a facility in contrast to a precision system which pinpoints the exact precise location of something down to a dot on the map. It uses motion sensors (accelerometers) and rotation sensors (gyroscopes) to continuously calculate the position, orientation, and the velocity.
  - The accuracy of these systems is not very accurate but it can be used for location tracking.
  - The major disadvantage of this method is accumulation of measurement error. One major advantage is that it can accomplish simple indoor localization at the lowest cost [21].
  - They are ubiquitous in both healthcare and manufacturing, as well as some other industries [22].
- **Radio based systems**

    - Radio based positioning systems transmits radio signals using beacons or tags (UWB, Bluetooth), picked up by the readers listening to these signals [30].
    - It has an accuracy of room level, but radio waves can sometimes be picked up by other readers through walls which acts as a disadvantage for RF based systems [23].
    - It requires extra hardware on different user devices therefore incorporating high cost of installation and use.
    - Radio technologies work better in open spaces characteristic of warehouses and manufacturing facilities.
- **Wi-Fi based systems**
    - In a Wi-Fi based system, tags are Wi-Fi transmitters that send simple packets to a number of Wi-Fi access points in a facility. These access points report the time and strength of that reading to a backend, which uses algorithms to compute position.
    - They are widely available and has high accuracy from three to five meters [24].
    - It does not require complex extra hardware which is the reason the setup and installation process is inexpensive.
    - Wi-Fi based positioning systems are commonly used in both healthcare and manufacturing settings.
- **Acoustic systems**
    - A number of new indoor positioning systems have come onto the market that use ultrasonic sound pulses from tags to locate them within an indoor environment. The tags emit a sound in the ultrasonic range. Receivers in the room (sometimes multiple, and sometimes a single "smart" one) pick up those sounds and locate the tags that way [25].
    - It has got high accuracy in locating users but is prone to sound pollution.
    - The cost of these systems could be high (for high end sonar based systems), and would require extra hardware depending on the infrastructure.
    - For now, acoustic systems are an uncommonly-used, but it has a good future with proprietary applications.
- **Vision based systems**
    - Infrared-based indoor localization systems use infrared light pulses (like a TV remote) to locate signals inside a building. IR receivers are installed in every room, and when the IR tag pulses, it is read by the IR receiver device. Just like the infrared systems we have LED based systems [26].
    - It is a fool proof way to guarantee room-level accuracy.
    - While the tags are low-cost and long-lasting, a drawback of infrared is that every room needs a wired IR reader to be installed in the ceiling which is expensive. In comparison

LED based systems or the camera based devices could be very expensive to setup.

- These are commonly used in construction, where rooms are definitively segmented. In an open-space warehouse, infrared would be a challenge. The most recent technology in this category is the LED-based system which has been adopted by Philips for the French supermarkets in 2015 [27].

Just like GPS there are many use cases where IPS can be leveraged – everywhere from corporate office campuses to shopping malls and airports.

Senion [29] a company deploying indoor positioning solutions worldwide mentions some of the examples where IPS is making a difference:

- **Mall of America** rolled out Indoor Positioning in 2017 to allow its visitors to find their way around the complex. MOA allows a rough estimate of 40 million annual visitors to use their localisation technique which offers step-by-step wayfinding, a way for guests to more easily connect with their brands and attractions.
- **Ericsson's** HQ in Stockholm which is catering around 4000 employees scattered across several buildings was facing a major challenge of an employee or group finding a certain meeting room at their HQ. To help employees save from this hassle, they decided to integrate Indoor Positioning. Their initial roll-out focuses on productivity improvements by saving employees time they might waste searching for things, such as conference rooms or places to work.

Some other industries successfully using indoor positioning includes [28]:

- Accessibility aids for the visually impaired
- School campus
- Museum guided tours
- Shopping malls, including hypermarkets
- Store navigation
- Warehouses
- Factory
- Airports, bus, train and subway stations
- Parking lots, including these in hypermarkets
- Hospitals
- Hotels
- Sports
- Indoor robotics
- Tourism
- Amusement Parks

1.2 Background of indoor positioning systems

The success of outdoor positioning and applications based on the global positioning system (GPS) provides an incentive to the research and development of indoor positioning systems. Unfortunately, the GPS system cannot be used effectively inside buildings and in dense urban areas due to its weak signal reception when there are no lines-of-sight from a Mobile-Station (MS) to at least three GPS satellites [8]. As a result, indoor positioning systems require alternative means to detect the MS's location without relying on the direct radio frequency (RF) signal from GPS satellites. As stated earlier- Infrared, RF, and ultrasound signals are major technologies used for indoor positioning systems. Different types of sensors are required to detect these electromagnetic signals which have characteristics depending on each location. For instance, a photo-diode detector is commonly used as a sensor to detect infrared signals. A sensing process converts these signals into a measurable metric such as distance or angle for later location determination [3]. Then, the measurable metrics are processed by a positioning algorithm to estimate the MS's position [3]. Unlike outdoor areas, the indoor environment imposes different challenges on location discovery due to the dense multipath effect and building material dependent propagation effect. Thus, an in-depth understanding of indoor radio propagation for positioning is crucial for efficient design and deployment.

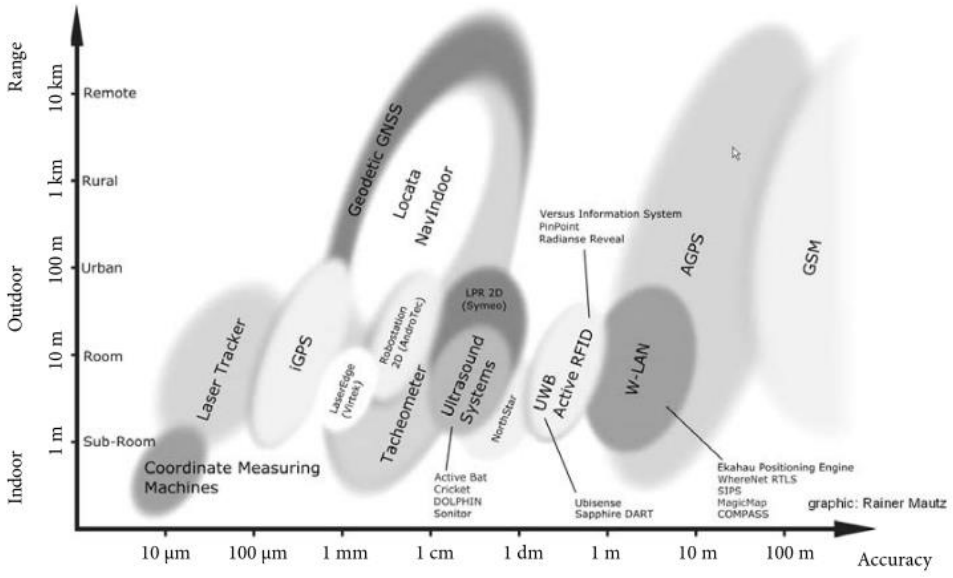


Fig. 1.2: Positioning systems according to their accuracy and coverage area [31]

Concurrently, there has been an increasing deployment of wireless local area networks (WLANs) by many individuals and organizations inside their homes, offices, buildings, and campuses. The popularity of WLANs opens a new opportunity for location-based services. The WLAN infrastructure can be applied to provide indoor location service without deploying additional equipment [9]. A

wireless network interface card which has the ability to measure RF signals can be considered as a kind of sensor device. Location-aware applications for indoor systems are potentially new emerging value-added services for WLANs and can possibly become a prevalent and successful technology of the future. Indoor positioning is an emerging technology that lacks theoretical and analytical background. Krishnamurthy [1] identifies three areas of challenges in position location in mobile environment which are performance, cost and complexity, and application requirements. These issues are summarized as follows:

- **Performance**: The most important performance metric is the accuracy of the location information. This is usually reported as an error distance between the estimated location and the actual mobile location. The report of accuracy should include the confidence interval or percentage of successful location detection which is called the location precision. Other essential performance metrics are delay, capacity, coverage, and scalability of the positioning system.
- **Cost and complexity**: The cost incurred by a positioning system can come from the cost of extra infrastructure, additional bandwidth, fault tolerance and reliability, and nature of deployed technology. The cost may include installation and survey time during the deployment period. If a positioning system can reuse an existing communication infrastructure, some part of infrastructure, equipment, and bandwidth can be saved. Trade-offs between the system complexity and the accuracy affects the overall cost of the system.
- **Application requirements**: The major application requirements for the location information are the granularity, performance, and availability. These requirements are different from one application to another. First, the granularity can be subdivided into temporal granularity and spatial granularity. Temporal granularity determines the rate at which the location information is requested while spatial granularity determines the level of detail of location information. Second, the performance requirements can include any combination of performance metrics discussed above. Finally, based on the type of applications, the location information may be required at different entities within a wireless network either at the mobile station itself or at a node within the backhaul network.

1.3 Future aspects

In practice, the evolution of the underlying technologies has had a very positive impact on the evolution of indoor positioning systems. The organisations have realized that changes in the subsequent versions of standards in a given technology can reduce some tasks in the positioning system or even solve some limitations. The method, the technology, and also the implementation details affect the accuracy of the system.

Some of the following areas which can be explored using IPS in the future are discussed below:

- **Location based marketing and insights**: With an indoor positioning system you can also collect insights from heat mapping, so you can see how people move around your venue. This allows you to gain important insights on what's good about your layout and what could be improved to increase sales at merchandise stalls, minimize wait times at food and drink vendors, or differentiate stand pricing at conventions.
- **Smart offices**: In the modern office, agile work methods are becoming increasingly adopted. Indoor positioning is used to reduce friction and increase productivity by helping employees book and cancel resources, help colleagues find each other more easily, find their way to meeting rooms as well as measure occupancy. Combining location and calendar information, proactive assistance can be given to the employee about things that are about to happen.
- **Shopping complex**: Enable a personalized shopping experience in the visitors' smartphones. Provide dynamic indoor GPS wayfinding to the stores, right in the smartphone. Analyse the physical flow of shoppers through detailed statistics and heat maps.

Fig. 1.3: Indoor positioning application combining wayfinding and visitor analytics [29].

1.4 Assumptions and contributions

There is not yet an overall satisfying solution for the IPS problem. Either very precise solutions are very expensive, or not real time, or cheap proposals are too inaccurate. Various machine learning algorithms have been used to resolve the indoor positioning problem which has resulted in different accuracies and processing time. This dissertation is a systematic study of the accuracy achieved by

various machine learning algorithms for wireless indoor positioning systems based on the location fingerprinting technique.

The overview of this dissertation is shown by the flowchart representing the main contributions in the figure below. Beginning with an investigation of the properties of the received signal strength of the WAPs which involves signals pre-processing. Next, a hybrid model of the positioning system is proposed consisting of two main components: the k-means clustering model for segregating the training and test sets into smaller clusters and the KNN positioning algorithm which is applied to each cluster separately and the value of the user's location is then predicted. This resulting model is considered as an enhancement to the better performing KNN for the Wi-Fi positioning system. Finally, given a performance goal, the proposed model can be used to determine the user's location with a much better accuracy and less positional error.

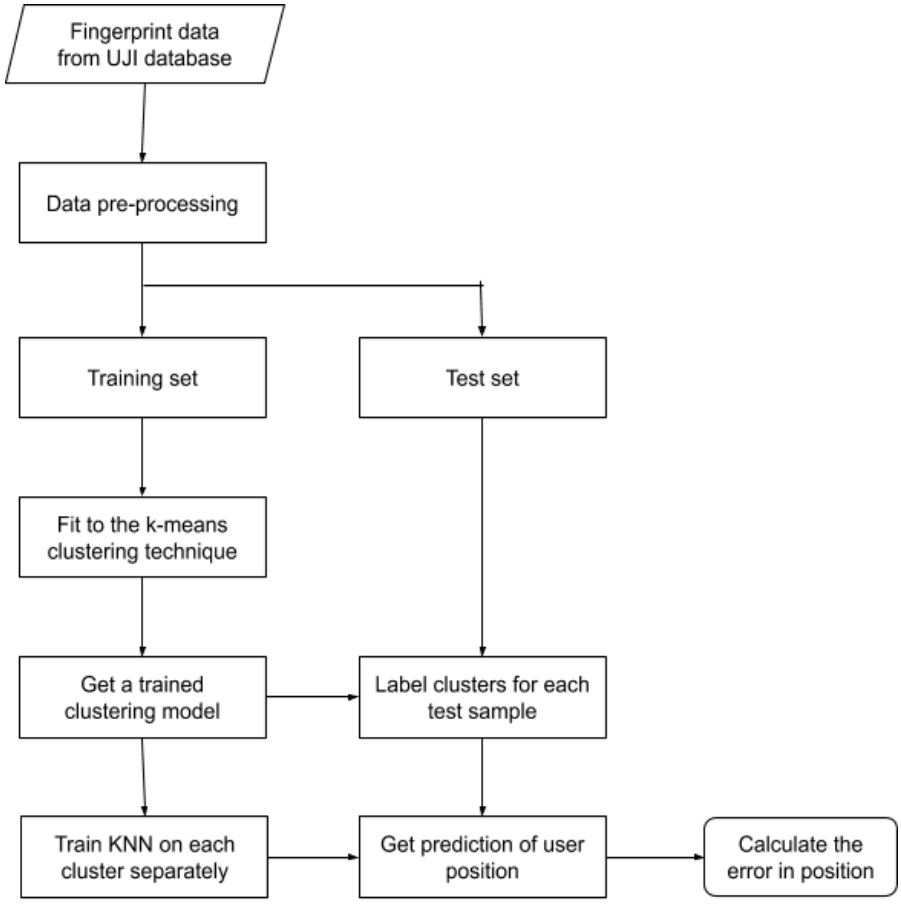


Fig. 1.4: Flowchart indicating the steps performed for our proposed hybrid model

There are five main assumptions that limit the scope of this work:

- This study is limited to the investigation of stationary mobile devices. No mobility tracking is considered.

- The placement of WLAN's infrastructure is not considered. The indoor positioning is assumed to be overlaid on top of existing infrastructure. Therefore, the performance of the positioning system depends on the placement of WLAN infrastructure.
- The optimum placement of WLAN's access points to support indoor positioning is not included in the scope of the current study.
- This study does not consider the search for an optimal distance calculation technique, but assumes generic algorithms as baseline (such as the Euclidean distance technique).
- Hybrid approaches that combine multiple sensor technologies is beyond the scope of this dissertation.

Starting with an analysis of the WAPs received signal strength from measurement experiments, this study performs an extensive data analysis of the location fingerprint in order to understand its underlying features. Properties of the location fingerprints are investigated in detail. In particular, it's statistics, and the distribution of the RSS is considered (whether it can be approximated by a Gaussian or lognormal distribution).

Currently, there are no clear guidelines on how to choose the minimum distance between physical positions. Moreover, it is not clear how many access points need to be "heard" at a given location for a given accuracy. The main goal is to study the accuracy and the precision performance metrics and suggest a performance evaluation methodology. The following is the list of contributions:

- Study and characterization of the unique properties of the received signal strength pattern in location fingerprints of the UJI database.
- Proposing a new hybrid algorithm to supersede existing positioning algorithms.
- Proposed a mathematical model for performance analysis of our hybrid algorithm i.e. calculating a combined Euclidean distance of predicted latitude and longitude from their true values.
- Developed a prototype of a software-based indoor positioning system to validate the proposed model.

1.5 Organisation of my dissertation

This dissertation is organized as follows:

Chapter 2, reviews the indoor positioning system and provides the justification of the direction of this dissertation.

In Chapter 3, the concept of indoor positioning with Wi-Fi is presented. It lists out the steps involved for data collection and the location estimation. It also mentions the machine learning models which could be used for solving the Wi-Fi fingerprinting problem.

Chapter 4 includes the analysis done on the competition dataset-mentioning the step by step process for our analysis, dataset metadata and the empirical results observed for the trained ML models.

Chapter 5 summarizes the results and suggests guidelines for future work along with the final conclusion deduced from the results. We also discuss the

challenges encountered and the things learned while working on the project as a self-assessment subsection.

2 Literature review

This chapter reviews the literature on wireless indoor positioning systems, as a means of providing an intellectual background for the present research. First in Section 2.1, the common components of indoor positioning systems are described. Then in Section 2.2, related indoor positioning systems that employ different technologies and techniques are briefly discussed besides radio frequency based WLANs. Finally, Section 2.3 and 2.4 reviews all relevant literature of indoor positioning systems by comparing the currently available systems and identifying their limitations.

2.1 Components of indoor positioning systems

A basic functional block diagram of wireless positioning system is suggested by Pahlavan et al [3]. It consists of a number of location sensing devices, a positioning algorithm, and a display system. Figure 2 from [3] illustrates these components and their relationships. First, the location sensing devices detect the signals transmitted by or received at known reference points using sensing technologies such as microwave radio frequency (RF), infrared, or ultrasound. The sensing technique – which can be based on time, direction (angle), frequency, or signal strength level – converts the sensed signal into location metrics that are time of arrival (TOA), angle of arrival (AOA), carrier signal phase of arrival (POA), or received-signal-strength (RSS) [3]. Given a set of known reference points, the relative position of the mobile station can be derived from the distance or the direction of these location metrics. Alternatively, the signal characteristics such as RSS at a particular location can form a pattern unique to that location. Then, the positioning algorithm processes the location metrics and estimates the location information using approaches such as signal processing [2], distance based approach [9], neural networks [12], or probabilistic approach [13]. Finally, the display system converts the location information into a suitable format for the end user.

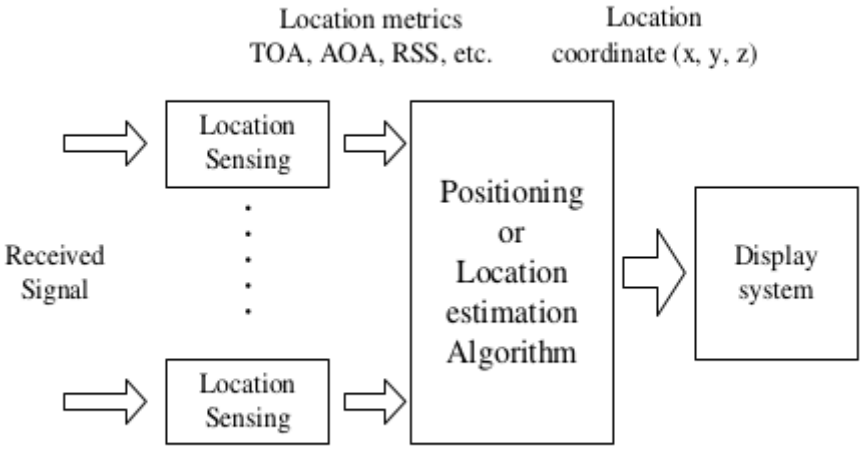


Fig. 2.1: A functional block diagram of positioning systems

Alternatively, a location system can be viewed from a software engineering perspective using a location stack (analogous to the OSI protocol stack) proposed by Hightower et al [14]. The location stack framework is a layered software engineering model that divides the positioning problem into smaller research problems. It aims to facilitate the development of future ubiquitous computing systems using the location information. The location stack extracted from [14] in Table 1 is designed based on the properties of positioning systems which are the fundamental measurement types, the measurement combination approaches, the object relationship queries, the preservation of uncertainty, and the application of user's activities. However, this abstract model is in the early stage and does not have any interface specification between layers yet. Table 1 summarizes the description of each layer. Detailed descriptions can be found in [14]. Based on this protocol stack, this dissertation focuses on the second layer.

Table 2.1: Summary of a positioning system stack

Layer	**Description**
6. Activities	A system, such as a machine learning system, for categorizing all available context information including location into activities. Activities are semantic states defined by a given ubiquitous computing application.
5. Context fusion	A system for merging location data with other non-location contextual information such as personal data, colour, temperature, light level, and so forth.
4. Arrangements	An engine for probabilistically reasoning about the relationships (e.g. proximity, containment, geometric formations) between two or more objects.
3. Fusion	A general method of continually merging streams of measurements into a time-stamped probabilistic representation of the positions and orientations of objects. Through measurement fusion, differing capabilities, redundancies, and contradictions are exploited to reduce uncertainty.
2. Measurements	Algorithms to transcribe raw sensor data into the canonical measurement types along with an uncertainty representation based on a model of the sensor that created it.
1. Sensor	Sensor hardware and software drivers for detecting a variety of physical and logical phenomena.

As discussed in Chapter 1, indoor positioning systems can be categorized based on their sensing technologies, measurement techniques, or system properties. The sensing technologies refer to the types of signals used by sensors, while the measurement techniques refers to the methods and metrics used in location sensing. Alternatively, Hightower and Borreiello [4] suggest a taxonomy of positioning systems based on system properties that are independent of sensing techniques and

measurement technologies. Their taxonomy suggests a guideline for evaluating positioning systems; however, some properties are not applicable to all positioning systems.

The distance measurement technique is usually called lateration, while the angle measurement technique is usually called angulation. Both lateration and angulation are subcategories of triangulation [4] that utilizes triangle geometry in determining a location. Besides these major categories, proximity, scene analysis, and other non-geometric features such as light level or temperature can be used as metrics in location measurement [10]. For instance, "proximity" uses a known location close to the object to determine the location, while "scene analysis" infers the location based on passive observation of features of a scene. The distance measurement is the most frequently used metric for location estimation. It can be estimated from the attenuation of signal strength based on the path loss and the time of flight (ToF) of signal based on propagation speed.

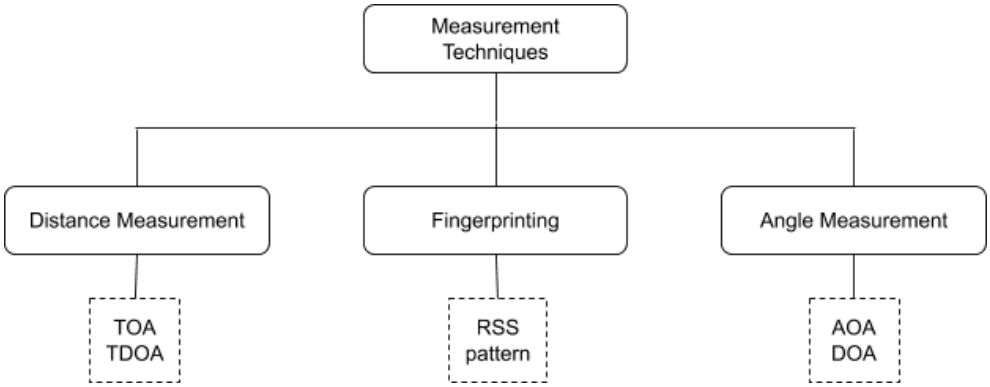


Fig. 2.2: Types of measurement techniques for indoor positioning systems

The first two techniques have been studied extensively for outdoor positioning systems [2]. They are suitable for systems with direct line-of-sight, but have problems or require complex computation in radio channels with noise, interference, and multipath. In indoor environments, the mobile station is surrounded by scattering objects which results in multiple angles of the signal reception. On the other hand, the distance between transmitter and receiver is usually shorter than the time resolution that can be measured by the system. Therefore, the AOA and TDOA approaches are impractical for indoor environments. The fingerprinting technique has gained more attention lately due to its simplicity compared to the first two for indoor positioning systems.

2.2 Related indoor positioning solutions

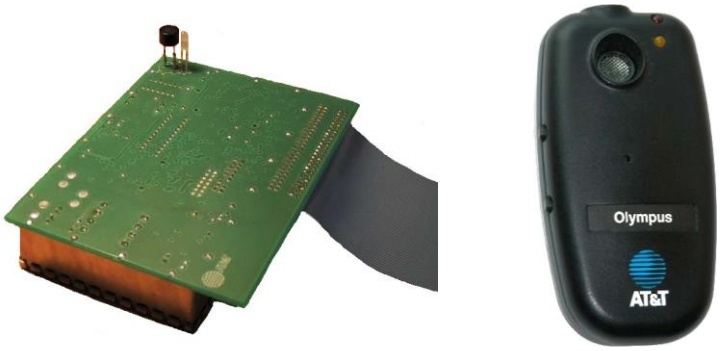
There are many indoor positioning systems with different architectures to determine the location of objects. They have different accuracies, configurations, and reliabilities. Some of the outstanding IPS are AT&T Cambridge Ultrasonic Bats, Microsoft Research's WaveLAN system, Active Badges, Smart Floor from Georgia Tech, Radio tags, and Computer vision systems. An excellent comprehensive

surveys of positioning systems can be found in [4] and with a special focus on indoor positioning systems in [10]. A subset of these systems is reviewed as examples below in which the major characteristics of these systems are summarized:

- **Active badge**: Active badges developed by AT&T Cambridge was the first indoor location sensing system [15]. The hardware consists of a small infrared beacon which is worn by every person in the complex, each beacon emits a unique code identifier after every 15 seconds. IR sensors are installed throughout the building. A network of these sensors detect the transmitted signals from the beacons. The data from these sensors is then collected by a central server which then stores it into a data bank and determines the location of the badge or the wearer.

Fig. 2.3: This picture shows four generations of the Active Badge. Bottom left, the first version, with a unique five bit code. Bottom right, the second version, with a ten bit code. Top left the third, current, version, with a forty-eight bit code, bi-directional capabilities, and an on-board 87C751 microprocessor [15]

- **Active bats**: A better and more accurate indoor positioning system has been developed by AT&T Cambridge which uses ultrasonic tracking technology [16]. Bats are the ultrasonic tags. The users and objects are tagged with these bats. These bats emitted periodic ultrasonic signals to receivers mounted across the ceiling. But there is a problem of using this ultrasonic technique as it requires a large number of receivers across the ceilings and a sensitive alignments for their placements across the ceiling.



(a) Bat Beacon Receiver (b) Bat Tag

Fig. 2.4: Components of the Active Bat ultrasonic positioning system. On the left is the receiver called Bat Beacon and on the right is the Bat Tag [16]

- **Cricket**: MIT Laboratories has developed small ultrasonic devices called the Cricket nodes. Figure below shows a Cricket unit which is a transmitter and a receiver application board. This system provides a positioning accuracy of 1–2 cm in an indoor environment of 10m2 [17]. The Cricket unit can be programmed either as a beacon or a listener. It can also be used for real-time tracking.

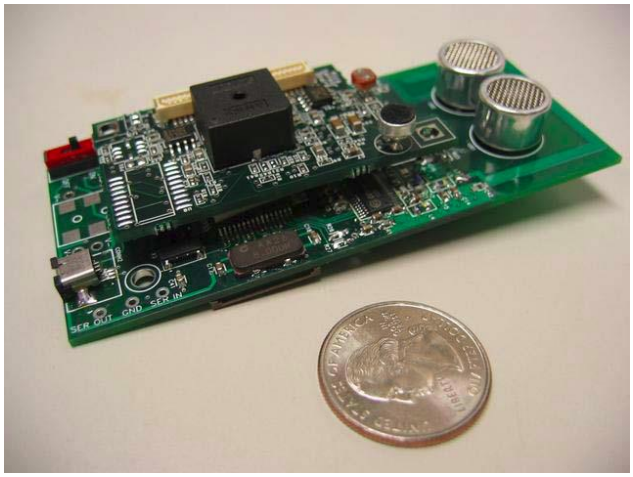


Fig. 2.5: The Cricket beacon and listener which are identical hardware devices [17]

- **DOLPHIN**: "Distributed Object Locating System for Physical space Inter networking" (DOLPHIN) [20] is an ultrasonic positioning system. It consists of distributed wireless sensor nodes. They send and receive the Radio and ultrasonic signals. It has an accuracy of 2 cm to be reached in a room of 3mx3m in size. These nodes are attached to various indoor objects. And using a novel distributed positioning algorithm in the nodes, DOLPHIN enables autonomous positioning of the objects with minimal manual configuration.

- **WaveLAN**: This system consists of a Network Interface Card (NIC) which provides the signal strength and signal to noise ratio which the WaveLAN system uses to determine 2D position of an object within a building by using the empirical data and applying it to the mathematical model [19].
- **LANDMARC**: LANDMARC uses RFID positioning system [18] where an RFID tag is preprogramed with an ID to be identified by the readers. RFID reader has the power range from 1-8 where level 1 is the shortest and level 8 is the longest range. The major advantage of LANDMARC is that it improves the overall accuracy of locating objects by utilizing the concept of reference tags.

Fig. 2.6: The RFID reader and tag used in our prototype system [18]

- **Computer vision**: Multi-camera systems can be used to track the people and to generate an intelligent environment around the complex. The system uses a set of stereo-coloured cameras to track multiple persons in the rooms. Stereo images helps to locate the person and colour images are used to maintain their identities. The system claims a location measurement accuracy of around 10 cm [10]. A disadvantage of the system is that uses multiple cameras to cover all the corners and as a result the hardware and setup cost is really expensive compared to the other systems.

Fig. 2.7: Digiclops colour 3D camera [10]

These pioneer works in this area have some disadvantages such as the limitation of the infrared or ultrasound sensing signals that cannot penetrate the

walls and floors which are common inside most buildings. The cost of sensor infrastructure installation and badges or tags for most of these systems becomes significant for a building with a lot of small rooms or offices.

2.3 Comparison of current solutions

A detailed comparison of the current wireless positioning systems is suggested by Hightower et al [14]. Table 2.2 summarizes the properties of these technologies. It consists of a number of location sensing devices and compares their system accuracy, its range, cost, the type of signal used and the principle.

Table 2.2: Comparison of the current sensing technologies

System	**Accuracy**	**Range**	**Signal**	**Cost**	**Measuring Principle**
Active Badge	7cm	5m	Infrared	Moderate	TOA, lateration
Active Bat	9cm	50m	Ultrasound	moderate	TOA, lateration
Cricket	2cm	10m	Ultrasound	low	TOA, lateration
Dolphin	2cm	Room scale	Ultrasound	moderate	TOA, lateration
Wave LAN	3m	Room scale	RF	moderate	RSS, triangulate
LandMarc	1-2m	50m	RF	moderate	RSS, triangulate
Computer Vision	10cm	Room scale	Camera Images	high	Image process

2.4 Limitations of current solutions

Almost all of the indoor positioning systems relies on the additional piece of hardware to be installed in the building. Sometimes this hardware is specially designed to cater the needs of the user. These hardware could be very expensive to install and maintain. Along with the cost there could be other limitations incurred during the process, some are listed below:

- **Hardware investment**: There's no doubt that indoor positioning is an investment and it's time consuming to set up. Setting up an indoor positioning system usually means deploying Bluetooth or RFID beacons, but it also means fingerprinting your complex, which is a time consuming task. The facts in literature indicates that buildings need at least 1 beacon per 200 m^2 in average though it is dependent on the physics of the building as well as the accuracy needed [14].
- **Maintenance**: Another disadvantage of IPS is the maintenance. The beacons themselves don't require other maintenance than a new

battery once in a while, and the best BLE beacons actually run five to eight years before their batteries need to be changed. However, if you have 500 beacons, changing the batteries may take a while. Add to that the work it requires to add more beacons if your venue expands.

- **Accuracy**: How accurate do you need the positioning to be? It's a given you want a reliable solution, but if you're guiding travellers around your airport, you don't need the same accuracy as if you're guiding people to the right book at a library or shoppers to the item on the shelf. You also need to think about the power consumption or if you need additional features like analytics, offline positioning, etc. Sensitive to light and noise [14].
- **Infrastructural changes**: Infrastructural changes would require remapping of the complex which would add to the installation and maintenance costs. We need to consider these factors before determining the type of technology required.

The above mentioned challenges are the main hurdles for these systems and as a result they are not used widely. For instance only 9 institutions in the world can afford the active bats [16].

Overall, the advantage of the infrastructure-free approach is the ease of deployment, at the expense of a lower positioning accuracy than the infrastructure-based approach.

3 Wi-Fi fingerprinting with machine learning

Wi-Fi positioning systems piggyback on top of the structures that already exist in the building (e.g. the Wi-Fi network) to provide the positioning service. The challenges for these systems are that the sensors in current smartphones are particularly noisy, because they are included for basic app support, rather than for the reliable positioning purpose. In this chapter we state how indoor positioning problem is solved by Wi-Fi fingerprinting in section 3.1. Section 3.2 gives a brief about the two modes of Wi-Fi fingerprinting technique, followed by the machine learning algorithms which could be deployed to predict users' location using Wi-Fi fingerprinting.

3.1 Indoor positioning with Wi-Fi fingerprinting

The proliferation of lightweight, portable computing devices and high-speed wireless local-area networks has enabled users to remain connected while moving about inside buildings. This emerging paradigm has generated a lot of interest in applications and services that are a function of a mobile user's physical location. The goal here is to enable the user to interact effectively with his or her physical surroundings.

Wi-Fi positioning is an indoor positioning system that uses the characteristics of nearby Wi-Fi hotspots and other wireless access points to discover where a device is located. It takes advantage of the rapid growth of wireless access points in urban areas.

Wi-Fi indoor positioning can be divided into two main categories. One of those categories relies on computing distances among handheld devices (such as a mobile phone) and the points whose coordinates are already known. This category takes advantage of the wave propagation technique. The second category is based on the radio map technology which combines the signal strength measurements and geographical coordinates. The most common and widespread localization technique used for positioning with wireless access points is based on measuring the intensity of the received signal (received signal strength indication or RSSI) called the method of "fingerprinting" [12]. In this dissertation the rest of the discussion focuses on the use of fingerprinting technique to solve the indoor positioning problem and how a machine learning algorithm could help us achieve that accurately.

This system is built on a deployment of off-the-shelf wireless LAN technology. Access points (or base stations) are located in such a way as to provide overlapping coverage in the area of interest. A mobile user carries with him/her a computing device equipped with a wireless LAN card capable of bi-directional communication with the access points [35]. Whole space where we want to use fingerprinting must be divided into smaller areas represented by fingerprints, which consists of RSS (Received Signal Strength) measured for each area separately. This data is then stored on a database. Of course, this requires an extensive mapping activity and storing Wi-Fi patterns for each mapped point in order to build the database. As a result the accuracy of the model depends on the number of nearby access points whose positions have been entered into the database [12].

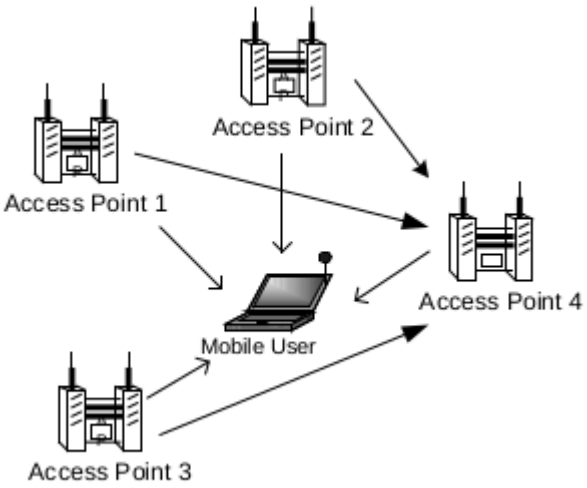


Fig. 3.1: Access point-based environmental profiling: Beacon packets from neighbouring APs are used to estimate (known) location of the target AP (AP4) using different Radio Maps [35]

In general, Wi-Fi radio map based positioning includes two phases: one is the primary phase is the training phase which is the primary phase where the wireless map of the complex is constructed using the field measurements and the secondary phase called the secondary phase called the positioning phase where position estimation is performed based on the wireless map. These phases are discussed further in the next section.

Compared with other positioning systems, Wi-Fi fingerprint positioning technology has the advantages of low-cost and high precision. Due to the wide deployment and use of Wi-Fi worldwide, fingerprint positioning technology can be applied to any indoor scenario where Wi-Fi networks are deployed without any additional hardware, which makes the technology cost low. This technology uses Wi-Fi signal strength to model and measure, without having to identify the exact location of the APs. With data networking speeds of up to 11 Mbps [36], wireless LANs have gained rapid acceptance and are widely being deployed in offices, schools, homes etc. Besides the existing wireless LAN our system does not require any additional hardware and can be enabled using purely software means.

Significant advantages of using Wi-Fi fingerprinting over other Indoor Positioning systems:

- Wi-Fi fingerprinting doesn't require Line-of-sight measurement of access points. The signals are the wireless signals which can penetrate through the walls, buildings or other obstructions with ease.
- There is no need of extra infrastructure to be installed in the complex. It makes use of the current available Wi-Fi access points and handheld mobile devices. Most of the public places and complexes have these already installed at their facilities.

Its main disadvantage is that any changes of the environment such as adding or removing furniture or buildings may change the "fingerprint" that

corresponds to each location, requiring an update to the fingerprint database. However, to deal with this situation i.e. the changing environment we can use it in integration with other sensor such as a camera.

3.2 Modes of Wi-Fi fingerprinting

As stated above the Wi-Fi fingerprinting is usually conducted in two phases: an offline phase often referred as survey phase followed by an online phase or the query phase. These phases are briefly explained below with the help of a diagram indicating the flow of information to and from the database and how it is used in the online phase to determine the location of a user.

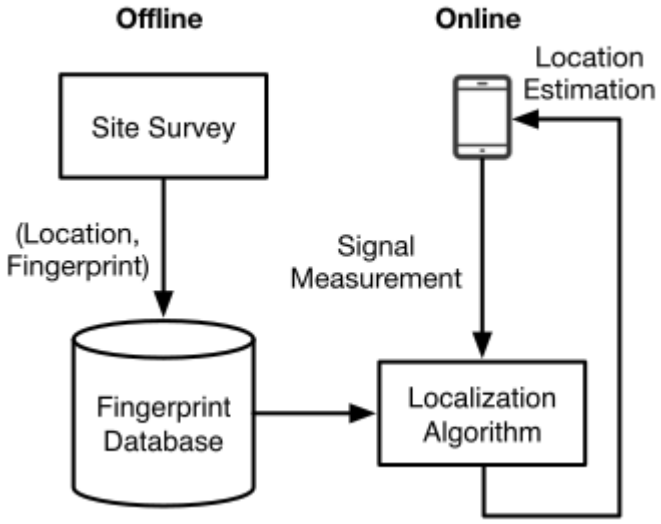


Fig. 3.2: Modes of Wi-Fi fingerprinting [35]

3.2.1 Offline mode

The fingerprint data collection is the training phase. This phase is an offline process and it needs to be redone if there have been major changes to the complex that could affect wireless propagation these could include relocation of access points, redesign of the infrastructure, etc.

The main task in this phase is to collect fingerprint data to build a model. There are two main methods of collecting fingerprint data, one is accurate modelling and the other is empirical modelling. Accurate modelling is a channel propagation model which only needs the signal strength of several important positions to generate the entire radio map by calculating and expanding the data with the channel propagation model. On the other hand for Empirical modelling one needs to move physically to every training location in the building and perform a Wi-Fi RSS measurement for building a signal strength map.

To cover the entire area, researchers distribute different numbers of AP and select various numbers of RP based on indoor environments. The most primitive way is by point-by-point manual calibration, which can achieve the greatest

accuracy. The target area is partitioned into numerous pieces, i.e., locations, and dedicated surveyors collect fingerprint samples point-by-point by considering the centre of each location as a measurement point.

This might accompany a Signal Pre-processing phase which may or may not be performed depending on the requirements, infrastructure and some other factors. The main purpose of signal pre-processing is to optimize the fingerprint database and remove invalid and redundant data by filtering and clustering, to improve the operating efficiency of the whole positioning system.

3.2.2 Online mode

In the online phase the positioning system match the pre-collected data with the signal strength at the user's location to determine where the user is. During this phase the use of positioning algorithms is important, there are two kinds of positioning algorithms, one is the deterministic positioning algorithm and the other one is the probabilistic positioning algorithm.

The deterministic positioning algorithm utilizes real-time matching of fingerprint data by means of algorithms such as machine learning techniques. This method is based on the measurement of signal strength; therefore, it is affected greatly by the precision of the fingerprint data acquired offline. Whereas the probabilistic positioning algorithm stores the probability distribution of the signal strength during a certain time in the fingerprint database and the probability position of the user's location is calculated by the Bayesian theory system based on signal strength. Nonetheless, irrespective of the algorithm used the system must be able to provide a location response from a request with observed data using the previous learning. This phase depends on the previous one, it means, the methods to be applied can be more or less complex and expensive in terms of time and resources depending on the kind and quantity of the stored offline data.

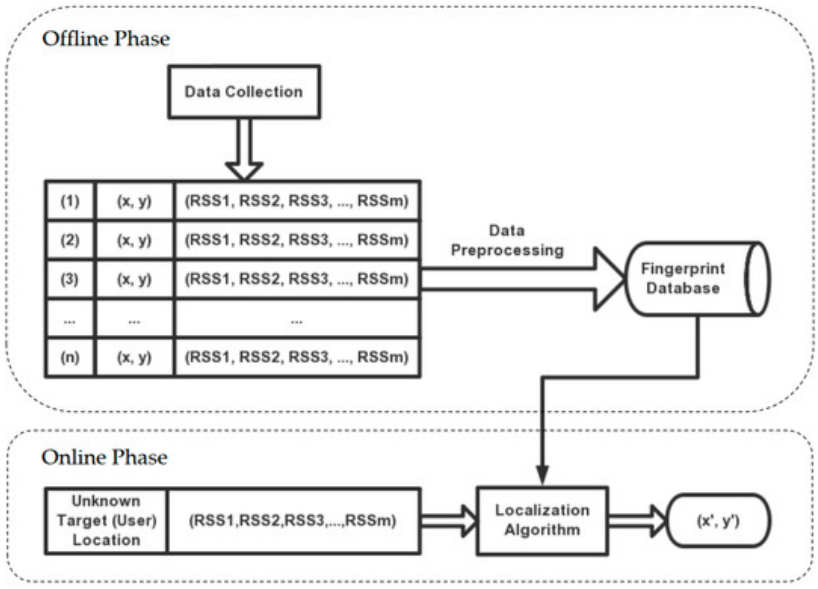


Fig. 3.3: Flow of data stored and retrieved in each phases of Wi-Fi fingerprinting [36]

The above figure illustrates the relation between both phases, showing the gathering of fingerprints on the offline phase in order to use it on the online phase through a localization algorithm when a user requires location estimation.

3.3 Machine learning and Wi-Fi fingerprinting

Location estimation algorithms or positioning algorithms are procedures that exploit dependency between the location information and its fingerprint in order to determine a position or location from samples of RSS signals. The examples of simple location estimation algorithms are strongest base station selection method and random selection method. The strongest base station selection assumes that the current user's position is closer to the base station that has the strongest signal strength, while the random selection reports the user's position at random from a set of known positions [32]. It is obvious that these two algorithms may not provide satisfactory results. More complex algorithms can take advantage of the dependency between RSS fingerprint and location information and could provide better accuracy, precision, and granularity of the location information.

From a machine learning perspective, the positioning algorithms estimate the location or position from samples of RSS vectors by learning from previous examples of location-dependent RSS fingerprints or signatures. The previous RSS data or training set are used to calibrate estimator models that can automatically relate location fingerprints and location information.

After a number of empirical and feasibility studies such as in [32, 33], recent development has been focused on the improvement of location estimation algorithms and system performance [5, 13, 12, 34]. Popular machine learning techniques such as KNN and neural networks have been introduced to improve the performance with RSS fingerprinting.

Wi-Fi fingerprinting could either be a classification or regression problem. If the aim is to detect the building or room (wider scope) then the positioning algorithms acts as a pattern classifiers, they are procedures that can automatically separate samples of patterns into different classes [13]. Each class is referred to as a class of RSS patterns that comes from the same location or position. It is considered a regression problem in which the aim is to predict user location coordinates. This dissertation predicts the physical coordinates thus considering it a regression problem. Supervised learning algorithms for regression are trained on data with the correct value given along with each variable. This allows the learner to build a model based on the attributes that best fits the correct value. By giving more data to the algorithm the model is able to improve. Learning can be described in this way as improving performance. The measure of performance is how well the algorithm predicts the regression value given a set of variables or attributes.

On the other hand unsupervised learning aims at finding the underlying structure of a dataset and to summarize it and group it most usefully. It is called "unsupervised" because we start with unlabelled data (there's no Y). Unlike supervised learning that tries to learn a function that will allow us to make predictions given some new unlabelled data, unsupervised learning tries to learn the basic structure of the data to give us more insight into the data.

3.4 Useful machine learning algorithms for Wi-Fi fingerprinting

In applications, predicting physical coordinates is treated as a regression problem. In this dissertation different machine learning algorithms are compared in terms of accuracy and processing time to determine the most suitable algorithm in indoor positioning.

Machine learning algorithms provide excellent solutions for building models that generalize well given large amounts of data with many attributes by discovering patterns and trends in the data; a task that is often difficult or impossible by other means.

With an increasing number of sensors being made available in the majority of mobile devices, large amounts of data can be collected and used to aid in the localization process. Machine learning algorithms are a natural solution for analysing through these large datasets and determining the important pieces of information for localization, building accurate models to predict an indoor position. Machine learning algorithms may also provide a fast, efficient method for indoor tracking, which will often be more useful to applications than static localization.

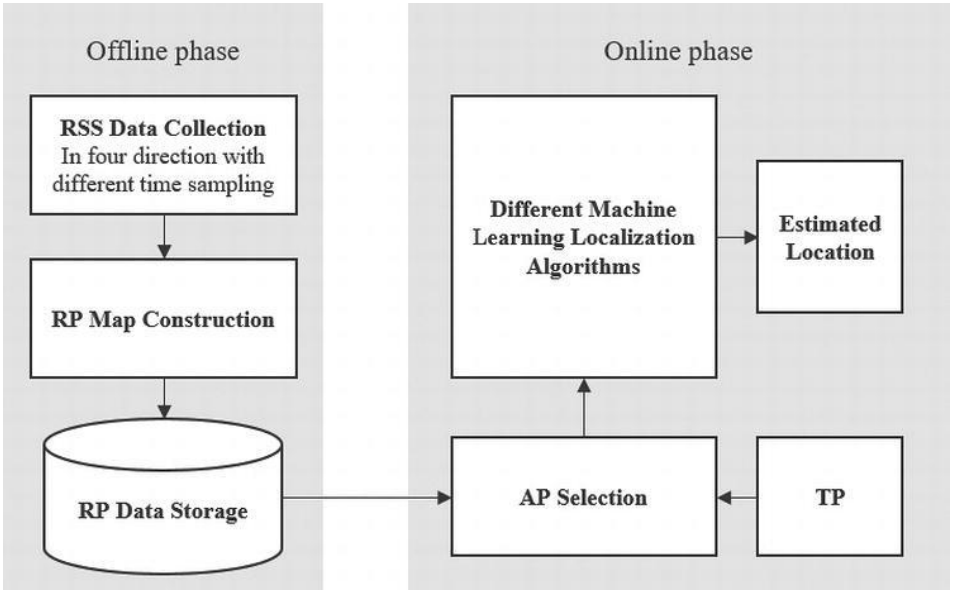


Fig. 3.4: Detailed Flow of data and information for indoor Wi-Fi localization using machine learning [37]

3.4.1 Linear regression

Linear Regression is usually the first and simple machine learning algorithm that could be used for a regression problem. It is a very powerful technique and can be used to understand the factors or features/attributes that influence the results or the target labels. It can be used to forecast sales in the coming months by analysing

the sales data for previous months. It can also be used to gain various insights about customer behaviour [38].

The objective of a linear regression model is to find a relationship between one or more features (independent variables) and a continuous target variable (dependent variable). When there is only one feature, it is called Univariate Linear Regression and if there are multiple features, it is called Multiple Linear Regression [39]. Wi-Fi fingerprinting is a problem of a multi-variate linear regression.

The linear regression model can be represented by the following equation:

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n \qquad \text{(I)}$$

where,
Y is the predicted value
θ_0 is the bias term.
$\theta_1, \ldots, \theta_n$ are the model parameters
$x_1, x_2, \ldots, x_n$ are the feature values.

The above hypothesis can also be represented by

$$Y = \theta_T x \qquad \text{(II)}$$

where,
θ is the model's parameter vector including the bias term θ_0
x is the feature vector with $x_0 = 1$

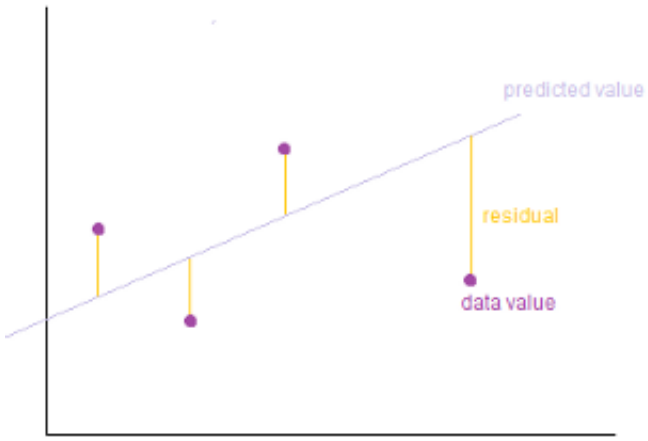


Fig. 3.5: Line of best fit to a multivariate datasets

The line for which the error between the predicted values and the observed values is minimum is called the best fit line or the regression line. These errors are also called as residuals shown in the following figure. To define and measure the error of our model we define the cost function as the sum of the squares of the residuals. The cost function is denoted by equation III:

$$J(\theta) = \sum_{i=1}^{M} (y_i - \sum_{j=0}^{p} \theta_j \times x_{ij})^2 \qquad \text{(III)}$$

Our objective is to find the model parameters so that the cost function is minimum.

Advantages

- They are often easy to fit, because we need to estimate only a small number of parameters.
- It results in a very simple model which is easy to interpret and deduce.

Disadvantages

- They make strong assumptions about the form of $\hat{y} = f(x)$. The parametric method will perform poorly if the specified functional form is doesn't represent the data correctly or the data doesn't fit to a linear line.

To avoid overfitting of data to the model we use regularisation techniques called Ridge and Lasso regression which are briefly discussed below:

- **Ridge regression**

Least squares regression isn't defined at all when the number of predictors exceeds the number of observations; It doesn't differentiate "important" from "less-important" predictors in a model, so it includes all of them. This leads to overfitting a model and failure to find unique solutions. Least squares also has issues dealing with multi-collinearity in data. Ridge regression avoids all of these problems.

In Ridge regression, magnitude of the coefficients is made as small as possible meaning each feature should have as little effect on the outcome as possible. This translates to having a small slope, while still predicting well on the data.

The goal of the algorithm is to minimize the cost function in equation IV:

$$J(\theta) = \sum_{i=1}^{M}(y_i - \sum_{j=0}^{p}\theta_j \times x_{ij})^2 + \lambda \sum_{j=0}^{p}\|\theta_j\|^2 \quad \text{(IV)}$$

A tuning parameter (λ) controls the strength of the penalty term. When $\lambda = 0$, ridge regression equals least squares regression. If $\lambda = \infty$, all coefficients are shrunk to zero. The ideal penalty is therefore somewhere in between 0 and ∞ [40].

- **Lasso regression**

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). The acronym "LASSO" stands for Least Absolute Shrinkage and Selection Operator [41].

Lasso regression performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients. This type of regularization can result in sparse models with few coefficients; some coefficients can become zero and eliminated from the model. Larger

penalties result in coefficient values closer to zero, which is ideal for producing simpler models.

The goal of the algorithm is to minimize the cost function in equation V:

$$J(\theta) = \sum_{i=1}^{M}(y_i - \sum_{j=0}^{p}\theta_j \times x_{ij})^2 + \lambda \sum_{j=0}^{p}|\theta_j| \quad \text{(V)}$$

Some of the θ's are shrunk to exactly zero, resulting in a regression model that's easier to interpret.

A tuning parameter, λ controls the strength of the L1 penalty. λ is basically the amount of shrinkage [41]:

- When $\lambda = 0$, no parameters are eliminated. The estimate is equal to the one found with linear regression.
- As λ increases, more and more coefficients are set to zero and eliminated (theoretically, when $\lambda = \infty$, all coefficients are eliminated).
- As λ increases, bias increases.
- As λ decreases, variance increases.

3.4.2 K nearest neighbours (KNN)

Linear regression is a parametric method which always have a trade-off for overfitting the data associated with it. On the other hand, the non-parametric methods provide an alternative and more flexible approach. K-Nearest Neighbours is a non-parametric approach for supervised machine learning. KNN is the simplest algorithm for regression. The aim is to find the K nearest neighbours to the new sample and predict with the average of their labels.

The KNN algorithm is one of the simplest ways to estimate location; it depends upon the Euclidean distance to measure the similarity/dissimilarity between the offline and online phases. KNN and its variants have been widely used in indoor positioning for its low-cost and high performance. The primary aim of KNN is to compare the signal strength obtained by users with fingerprint data in the fingerprint database while positioning. It chooses the k nearest neighbours of fingerprint data. It completes the positioning operation by calculating the weighted average of the k fingerprint data. The Weighted K-Nearest Neighbour is a variant of KNN that adds distance as weight. Nearest Neighbour is another variant of KNN in which k is set to 1.

Mean distance error is typically used as the performance metric and is calculated as the average Euclidean distance between the actual location and the estimated location.

The radar system is the earliest indoor positioning system that applies the nearest neighbour methods. Bahl et al. [32] presented a new version called the Nearest Neighbour in Signal Space by improving the nearest neighbour methods. The core ideal of the NNSS was to match the nearest neighbour by calculating the distance between the signal strength received by users and the signal strength in the fingerprint database. Bahl [32] compared this method with the perceived similarity and channel model methods and found that the positioning accuracy of this method

could reach 2.94 m. It still has some issues in its practical applications, but has provided new thoughts in researching indoor positioning.

The best choice of k depends upon the data; generally it is indicated through previous research and applications that larger values of k reduces effect of the noise on the prediction, but make boundaries between classes less distinct [42]. A good k can be selected by various hyper-parameter optimization techniques mainly by cross-validation. The special case where the class is predicted to be the class of the closest training sample (i.e. when k = 1) is called the nearest neighbour algorithm. The accuracy of the k-NN algorithm can be severely degraded by the presence of noisy or irrelevant features, or if the feature scales are not consistent with their importance.

The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.

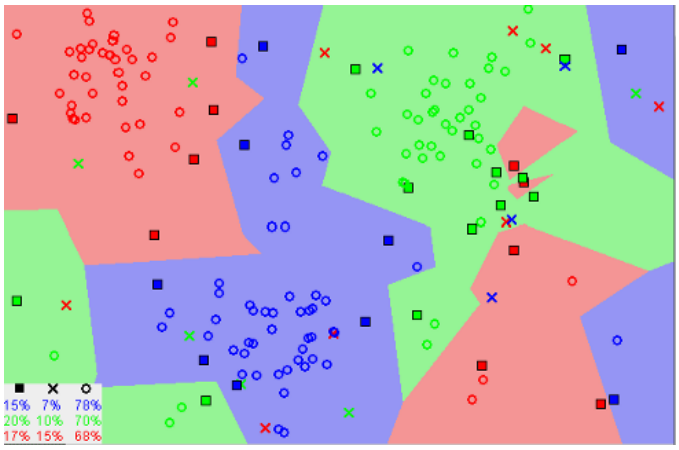


Fig. 3.6: Image showing how similar data points typically exist close to each other [43]

Notice in the image above that most of the time, similar data points are close to each other. The KNN algorithm focuses on this assumption being true enough for the algorithm to be useful. KNN captures the idea of similarity (sometimes called distance, proximity, or closeness) with some mathematics- calculating the distance between points on a graph.

Advantages

- The algorithm is simple and easy to implement.
- There's no need to build a model, tune several parameters, or make additional assumptions.
- The algorithm is versatile. It can be used for both classification and regression problems.

Disadvantages

- The algorithm gets significantly slower as the number of examples and/or predictors/independent variables increase.

However, KNN takes all the nearest K neighbours to calculate the estimated result, while not all of them do contribute. If we could select some of these K

neighbours before calculation, a more accurate estimate could be found. As shown in the figure below, if we could construct a sample group with respect to the grey area, we could get a much better estimate e2. So, we use clustering to filter out some of the useless neighbours.

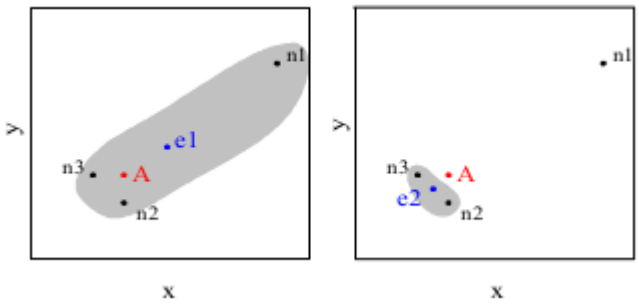


Fig. 3.7: Figure depicts that by selecting some of the K neighbours results in a more accurate estimate

3.4.3 K means clustering

K-means clustering is one of the simplest and most popular unsupervised machine learning algorithms. Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes. The objective of K-means is simple: group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (k) of clusters in a dataset. A cluster refers to a collection of data points aggregated together because of certain similarities.

We will first define a target number k, which refers to the number of centroids you need in the dataset. A centroid is the imaginary or real location representing the centre of the cluster. Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares. In other words, the K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. The 'means' in the K-means refers to averaging of the data; that is, finding the centroid.

To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids. It halts creating and optimizing clusters when either: The centroids have stabilized — there is no change in their values because the clustering has been successful. Or the defined number of iterations has been achieved [43].

The K-means clustering algorithm is used to find groups which have not been explicitly labelled in the data. This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets. Once the algorithm has been run and the groups are defined, any new data can be easily assigned to the correct group.

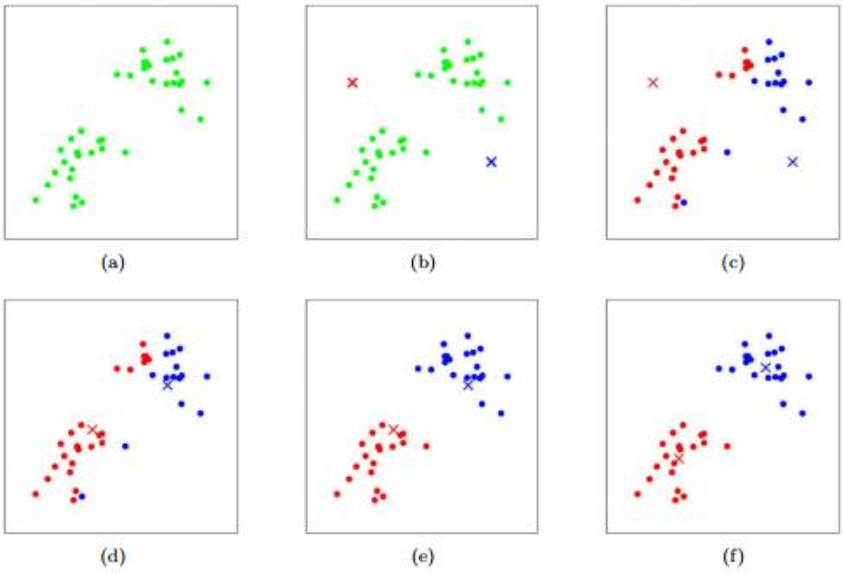


Fig. 3.8: K-means algorithm example. Training examples are shown as dots, and cluster centroids are shown as crosses. (a) Original dataset. (b) Random initial cluster centroids. (c-f) Illustration of running two iterations of k-means. In each iteration, we assign each training example to the closest cluster centroid (shown by "painting" the training examples the same colour as the cluster centroid to which is assigned); then we move each cluster centroid to the mean of the points assigned to it [44]

Ma et al. [45] presented a new method called the Clustering Filtered KNN (CFK) which combined clustering with the KNN method. CFK can divide the entire fingerprint database into several non-adjacent parts by using hierarchical clustering as per the physical location of each neighbour point. It selects the nearest neighbour points in one cluster to match the user's location. Ma also found that the average positioning error and positioning median error were both less than the KNN, as was the calculating cost.

4 Experiments and empirical observations

By analysing an objective database for comparing positioning systems & WLAN-fingerprinting algorithms, the main goal of our experiment is predicting people's location (coordinates) at Jaume I University from WAPs signal information using supervised regression (linear regression, lasso, KNN) and unsupervised machine learning algorithms (k-means clustering). The following section includes some brief information about the environment setup and the competition dataset used for our analysis. It further compares the empirical results obtained by using different machine learning algorithms in terms of their performance accuracy. A brief about the hardware setup required to run the analysis i.e the software proof-of-concept is stated in the Appendix-A.

4.1 Environment setup

Although there are many papers in the literature trying to solve the indoor localization problem using a WLAN fingerprint-based method, there still exists one important drawback in this field which is the lack of a common database for comparison purposes. So, UJIIndoorLoc database is presented to overcome this gap.

The UJIIndoorLoc database covers three buildings of Universitat Jaume I with four or more floors and almost 110,000 m^2. It was created in 2013 by means of more than 20 different users and 25 Android devices. The database consists of 19,937 training/reference records [46].

Unlike other databases, UJIIndoorLoc is considered superior for our analysis because of its following properties:

- The samples taken can be considered realistic as human users were used and not machine generated.
- A variety of mobile devices as well as a large number of users were used, thus widening the scope of our analysis.
- A large area was covered, more buildings with more than 1 floor were used and their internal structures differed.
- Validation samples have been provided, apart from the training samples.

This database can therefore be presented as a common, public database that can be used for comparisons.

4.1.1 UJIIndoorLoc testbed

UJIIndoorLoc, which can be downloaded from UCI Machine Learning Repository, is the most comprehensive indoor positioning database in the literature. Android apps were used to create the database. They provided geographic information of the interior of the buildings i.e. the training reference point's localization. – CaptureLoc for Training Data & ValidationLoc for Testing Data.

Main characteristics of the UJIIndoorLoc Testbed are:

- It covers a surface of 108,703m2, 2 – 3 buildings with 4 or 5 floors each.
- 933 different places (reference points) appeared in the database.

- 520 different WAPs (Wireless Access Points) were installed all around the buildings.
- 19,938 sample points were obtained for training/learning and 1,111 were obtained for validation/testing – 21,049 in total.
- Testing samples were taken 4 months after the training samples to ensure dataset independence.
- Data collected by more than 20 users with 25 different mobile device models.

4.2 Dataset description

The entire database is separated such that it contains 19,937 records reserved for training and 1,111 records for testing purposes. In total there are 529 features and these features are the coordinates where Wi-Fi fingerprints are taken, such as building, floor, space (office, lab, etc.), and relative position (in a room or corridor) etc. The following subsections gives a brief description about each attribute and the output labels that our algorithm will predict, followed by a 3d figure of each building included in our testbed which represents the coordinates along with the their floor number thus giving a rough idea of the structure and shape of the building.

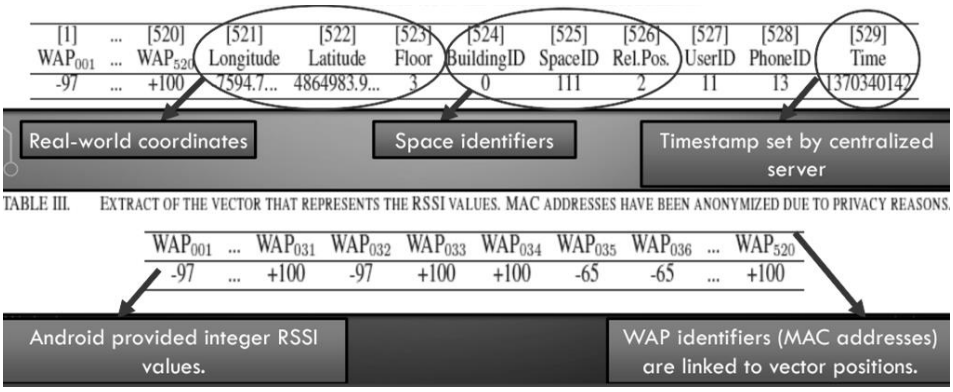


Fig. 4.1: UJIIndoorLoc database description with an example [49]

4.2.1 Attributes information and the output labels to predict

The 529 attributes contain the Wi-Fi fingerprint, the coordinates where it was taken, and other useful information.

Each Wi-Fi fingerprint can be characterized by the detected Wireless Access Points (WAPs) and the corresponding Received Signal Strength Intensity (RSSI). The intensity values are represented as negative integer values ranging -104dBm (extremely poor signal) to 0dbM. While, the positive value 100 is used to denote when a WAP was not detected. During the database creation, 520 different WAPs were detected. Thus, the Wi-Fi fingerprint is composed by 520 intensity values [46, 44].

The coordinates (latitude, longitude, and floor) and building ID are provided as the attributes to be predicted. The particular space (offices, labs, etc.) and the relative position (inside/outside the space) where the capture was taken

have been recorded. Outside means that the capture was taken in front of the door of the space. Information about who (user), how (android device & version) and when (timestamp) Wi-Fi capture was taken is also recorded [46].

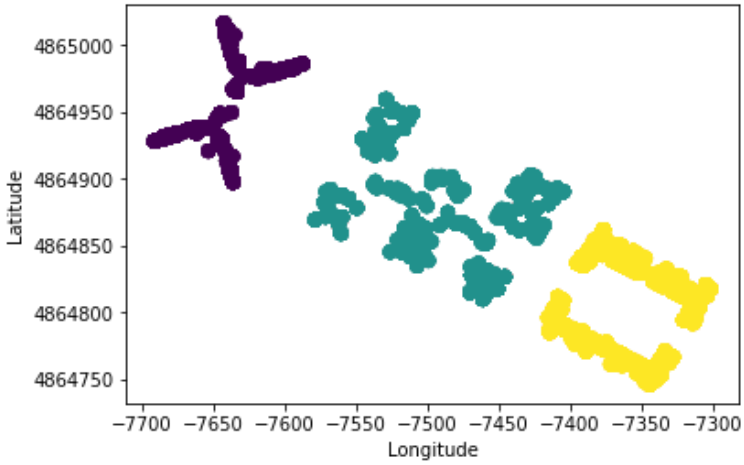
Brief description about the attributes [46]:

- Attribute 001 (**WAP001**): Intensity value for WAP001. Negative integer values from -104 to 0 and +100. Positive value 100 used if WAP001 was not detected.
  (...)
- Attribute 520 (**WAP520**): Intensity value for WAP520. Negative integer values from -104 to 0 and +100. Positive value 100 used if WAP520 was not detected.
- Attribute 521 (**Longitude**): Longitude. Negative real values from -7695.9387549299299000 to -7299.786516730871000.
- Attribute 522 (**Latitude**): Latitude. Positive real values from 4864745.7450159714 to 4865017.3646842018.
- Attribute 523 (Floor): Altitude in floors inside the building. Integer values from 0 to 4.
- Attribute 524 (**BuildingID**): ID to identify the building. Measures were taken in three different buildings. Categorical integer values from 0 to 2.
- Attribute 525 (**SpaceID**): Internal ID number to identify the Space (office, corridor, and classroom) where the capture was taken. Categorical integer values.
- Attribute 526 (**RelativePosition**): Relative position with respect to the Space (1 - Inside, 2 - Outside in Front of the door). Categorical integer values.
- Attribute 527 (**UserID**): User identifier (see below). Categorical integer values.
- Attribute 528 (**PhoneID**): Android device identifier (see below). Categorical integer values.
- Attribute 529 (**Timestamp**): UNIX Time when the capture was taken. Integer value.

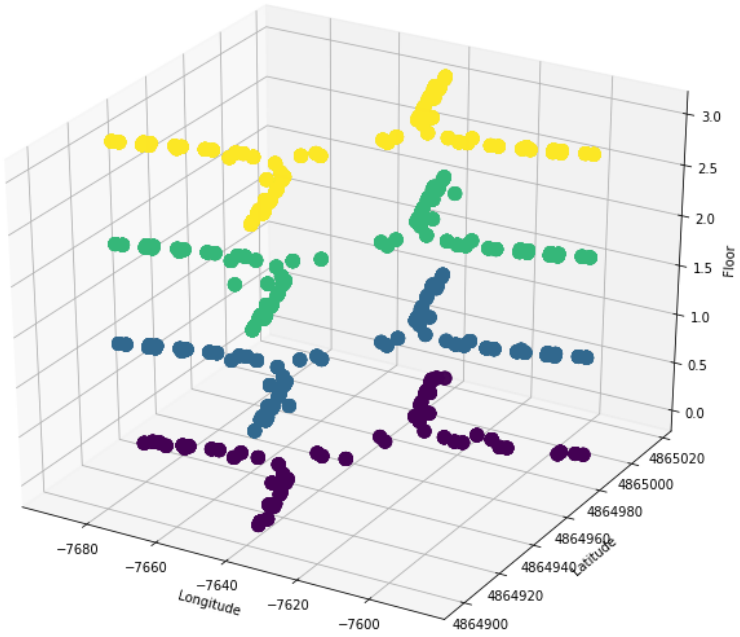
4.2.2 3D model of the testbed

Visualising is a powerful tool to get an initial rough idea about the problem. In this case we use clustering technique to divide the training samples into clusters and then visualise it on a 3D plot. Our k-means algorithm performs really well and is able to perfectly differentiate the samples into three different buildings as represented in the following graphs. We can clearly get an idea about the shapes of the buildings. Also we can recognise when looking in Google maps around the area, which building it is. It is part of the University Campus from Universitat Jaume I ,in

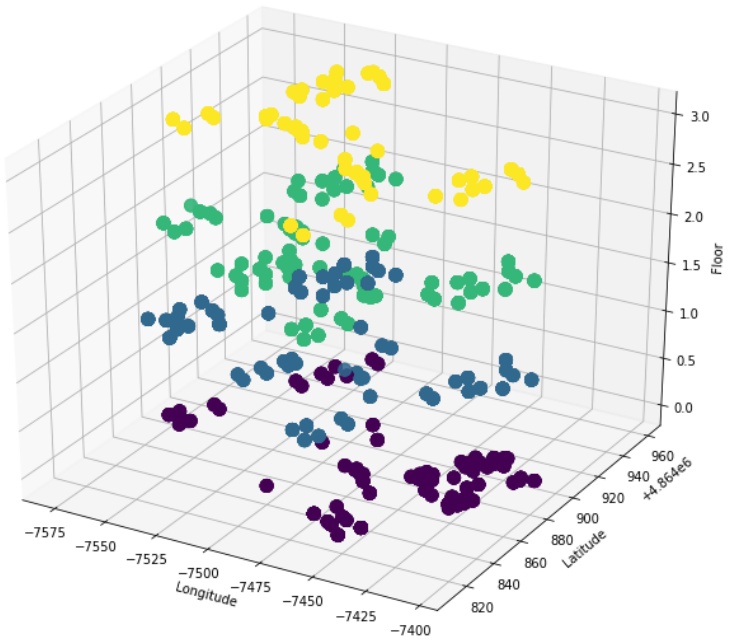
Castelló, Spain. The following plots shows the coherence of the data collected. It seems there is no mistake in its collection.



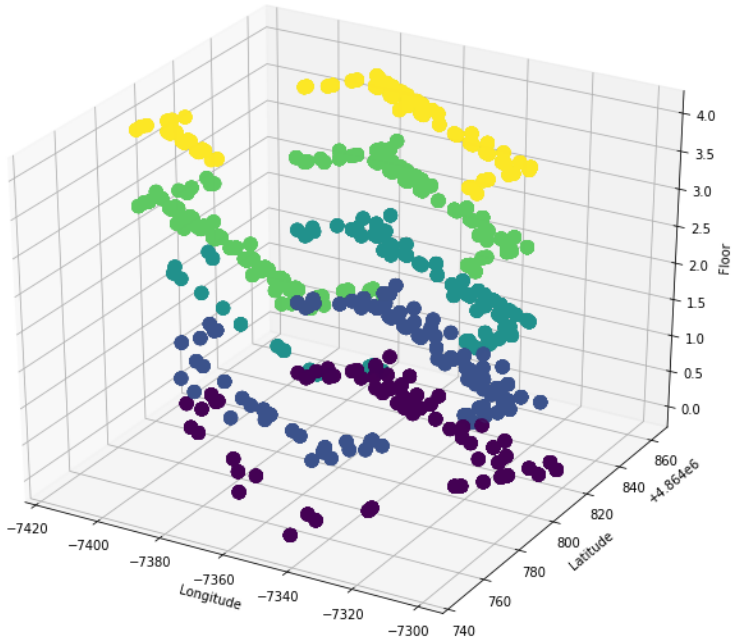
(a) 2D model of the testbed representing the three buildings



(b) 3D model of the first building with 4 floors



(c) 3D model of the second building with 4 floors



(d) 3D model of the third building with 5 floors

Fig. 4.2: Visualisation of the structure of our testbed using the training data samples.

4.3 Exploratory data analysis (EDA)

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. It is a good practice to understand the data first and try to gather as many insights from it. EDA is all about making sense of data in hand.

The following subsections mentions various EDA techniques that we use to get insights from our dataset.

4.3.1 Statistics of the dataset

Analysing the statistics of each attribute of our dataset is a typical quantitative technique used for EDA. In this study we first check the dimensions of the dataset and check for the null/missing values. We observe that there aren't any missing values in our dataset (neither in the train nor in test dataset).

Next, we calculate the general statistics of each attribute of our dataset which includes calculating the total count, minimum, maximum value for each attribute along with its mean, quartiles and standard deviation. These statistics gives us a deep insights about the dataset. It indicates some useful information about the attributes, including:

- There are some WAPs in the training dataset which contains only a single value i.e. 100db for the observed signal strength thus indicating that these WAPs are never detected. Therefore, it makes sense to remove these columns from our datasets. Values with variance of zero are only introducing noise in our model, therefore they are not useful.
- Some of the WAP attributes have a high value of standard deviation, thus indicating they are used quite often (this could be because these WAPs are located at the end of the building which is nearer to the other building thus can be detected).
- Original dataset used WAP signal range from -104 to 0 (-104 being the lowest signal and 0 the highest), and value 100 when WAP was not detected. We observe that 100 is the majority value for all of the WAPs attributes so for scaling purpose and to follow the standard we replace 100 with an arbitrary negative value lower than -104. We choose -115 as the value when a WAP is not detected.

4.3.2 Histograms representing the signal strengths

Histograms helps us identify the distribution of each of the feature. A distribution plot will help represent the spread of different values of data we have across our dataset and more importantly, help to identify potential outliers. Histograms depict the underlying frequency of a set of discrete or continuous data that are measured on an interval scale. This depiction makes it easy to visualize the underlying distribution of the dataset, and inspect for other properties such as skewness.

Traditionally, the average RSS is believed to be log-normally distributed according to the large-scale fading model [48]. The mean value is generally

predictable and believed to follow one of several standardized path loss models discussed in [23]. However, there are some conflicting conclusions regarding the RSS distribution measured at the software level by the wireless NIC for indoor radio propagation in [49] and [5]. Moreover, the standard deviation and the stationarity of the RSS are not understood very well.

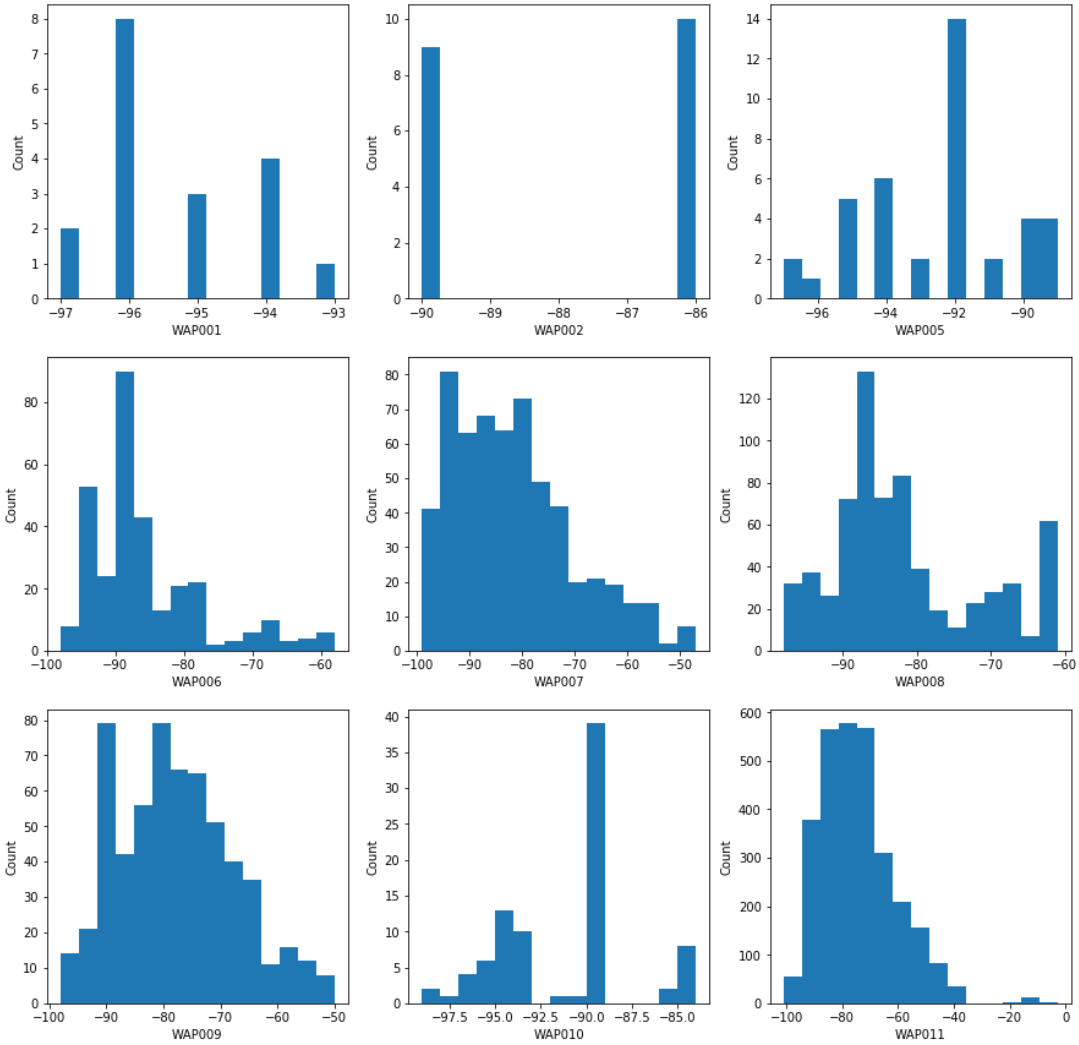


Fig. 4.3: Histogram distribution plots representing some of the WAPs in our dataset

By looking at the histograms we observe that most of the WAP attributes are left or positively skewed while a few exhibit a non-uniform behaviour throughout the training phase. Slightly skewed distributions often occur when the RSS level is low (the AP is far from the measurement location or there is no direct line-of-sight). These conditions are often valid for indoor environments.

4.4 Fitting the models

Now that we've cleaned and visualized the data, we can remove attributes that won't be needed in a predictive model. After this we will fit the models which we have implemented from scratch to our modified dataset.

As indicated in the last chapter, Wi-Fi fingerprinting could be solved by both classification as well as regression algorithms. A detailed explanation of the models used and their results are included in the following sections with the main focus on the regression algorithms used to predict the coordinates.

We start with a simple linear regression model for predicting the longitude and latitude coordinates of a user. As a measure of error, we would use Euclidean distance (equation VI) to calculate the error between the predicted and the true labels.

$$err = \sqrt[2]{(\hat{y}_{lg} - y_{lg})^2 + (\hat{y}_{lt} - y_{lt})^2} \qquad \text{(VI)}$$

where,
$\hat{y}_{lg}$ is the predicted value of user's longitude coordinate
y_{lg} is the true value of the user's longitude coordinate
$\hat{y}_{lt}$ is the predicted value of user's latitude coordinate
y_{lt} is the true value of the user's latitude coordinate

We will use L1 and L2 regularization techniques to improve our model's performance. On using Ridge regression there is a significant improvement in the performance. Better results are obtained with a Lasso model as indicated in the next section evaluating the performance of each model.

Next, we will perform KNN regression over the modified dataset with an initial default value of parameter k=1 (1 nearest neighbour). We will use the same error measurement technique for evaluating the performance of this model. We observe that there is a high decrease in the value of error compared to the linear models which indicates a possibility of overfitting in earlier cases.

Even though there is a significant decrease in the error obtained with the KNN model, there is still a chance to improve further by optimising the value of K. Now we need to find an optimal value for parameter K resulting in a better model. We will plot the error rates obtained for different values of k ranging from k=3 to k=49. It is observed that with k=5 we get the minimum total error on our test dataset. The results are compared in the next section.

The following figure represents the error curve obtained when the coordinates are predicted using different values of K. It can be concluded from the figure that the best results are obtained for K=5 after which the error gradually increases with increase in the value of K.

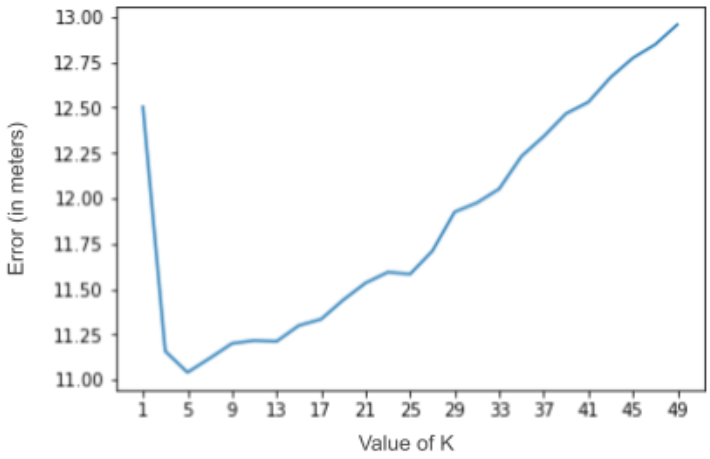


Fig. 4.4: Performance of KNN model with different values of K

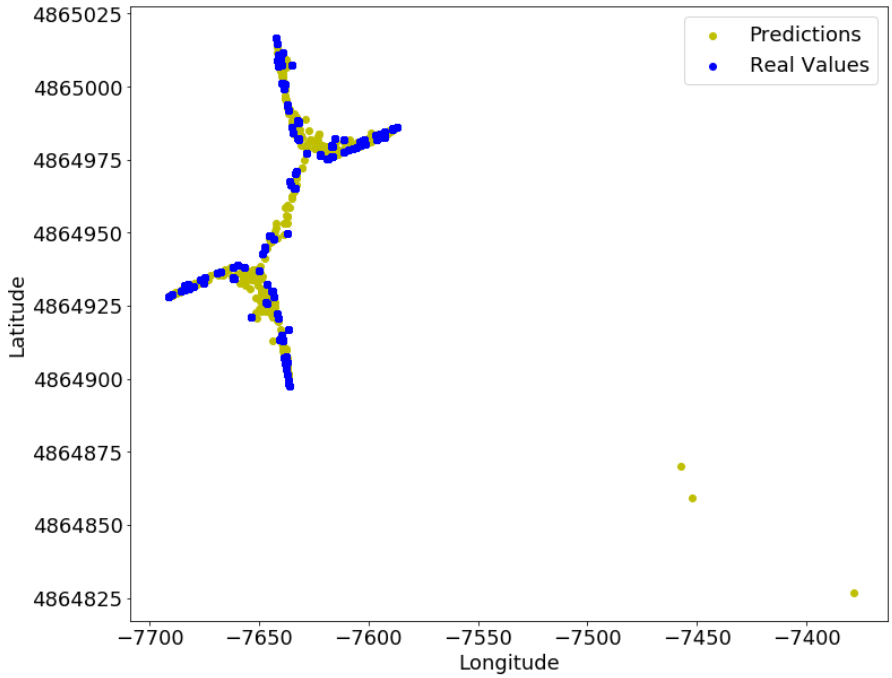
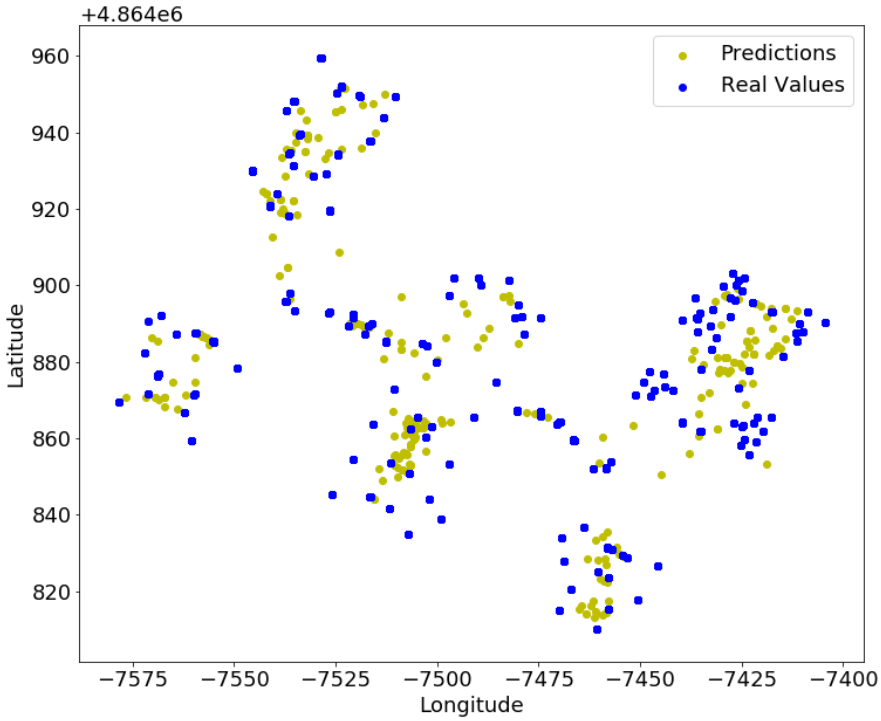


Fig. 4.5: (a) Predictions on building 1 using KNN with optimal value of K

From the plots in figure 4.5 we observe that even though the performance improved with the optimal value of K but still in some cases the model is labelling coordinates from another building thus resulting in wrong predictions. To counter

this a better approach would be if we can try segregate these samples into groups and then predict on individual groups using KNN.



(b) Predictions on building 2 using KNN with optimal value of K

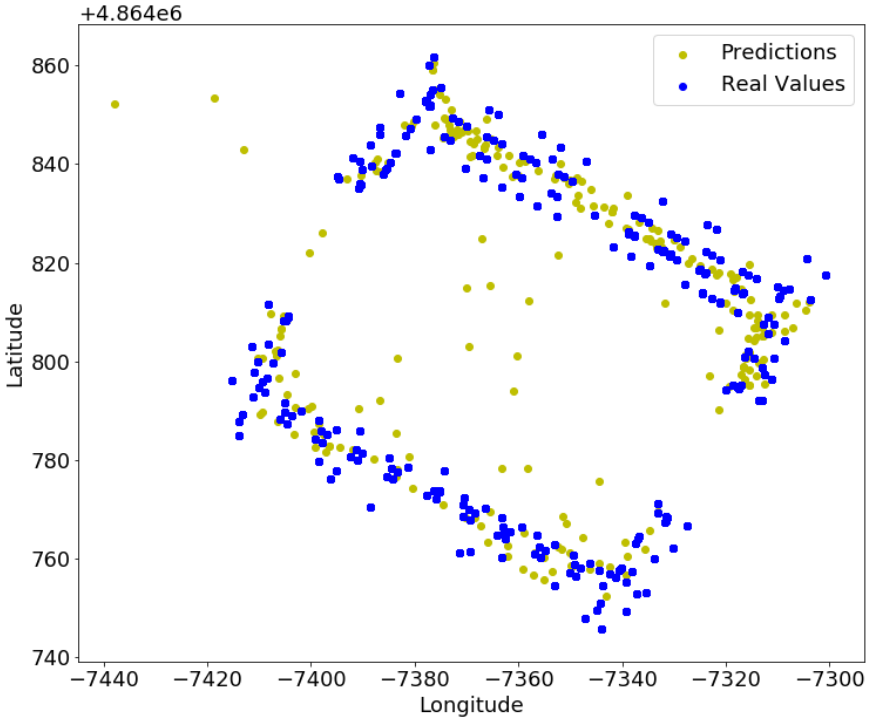


Fig. 4.5: (c) Predictions on building 3 using KNN with optimal value of K

So, finally we try to cluster the modified test dataset into different clusters using the k-means clustering model derived from the modified training data set. As a measure of accuracy we calculate the distance from the true coordinates to the centroids of all the three clusters and if the nearest cluster is the same as the predicted cluster then it is assumed to be the correct prediction. It is observed that our clustering algorithm does work really well with an accuracy of about 98%. This technique would result in categorization of the test samples into different clusters. We will then apply KNN to each cluster of test dataset and predict the coordinates.

We try our experiment with different values for the number of clusters and observe that as we increase the number of clusters the error decreases until it reaches a saturation value at n_clusters=49.

In contrast to the regression algorithms, classification algorithms also produce good results in predicting the floor and the building number. On our initial analysis to check for imbalance between the classes we construct the following pie-charts representing the percentage of train and test samples belonging to one of the 3 buildings.

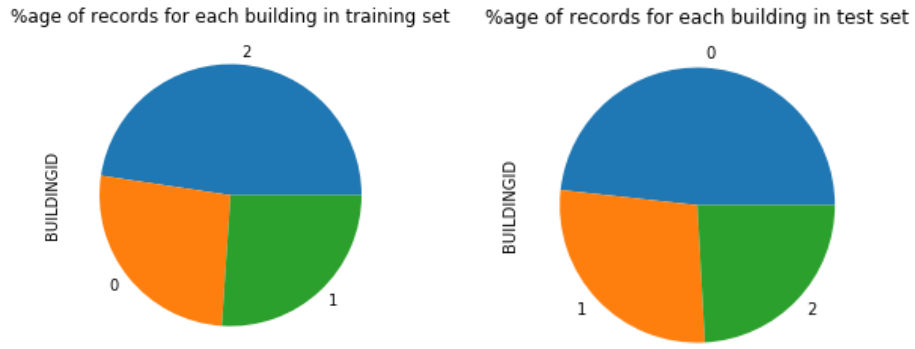


Fig. 4.6: Pie-chart representing %age of training (left) and test (right) samples belonging to one of the three building

We observe that there is a slight imbalance between the train and the test samples but not significant enough to affect our predictions. We use KNN model with K=5 for the classification of floor and building number and observe that there is a high accuracy of about 99% predicting the building number and 88% for the floor number.

Finally, it can be concluded that there is a significant improvement in performance observed in KNN with clustering compared to the results of KNN without clustering while predicting the coordinates as shown in the next section. The proposed algorithms evaluation will be presented in the subsequent subsections.

4.4.1 Performance evaluation

To evaluate the performance of the different machine learning techniques used in our analysis, the localization error was computed as the Euclidean distance

between the actual reported coordinates of the test points and the coordinates of the mobile user during the online phase.

We compare the localization accuracy of different deterministic (e.g. K-Nearest Neighbours (KNN), Ridge regression and KNN with clustering) algorithms in our experiments. Non-surprisingly, simple deterministic K-Nearest Neighbours (KNN) algorithm outperforms the Lasso regression algorithms in terms of localization accuracy. A better performance is obtained when KNN is combined with a k-means clustering technique. There is a slight improvement on increasing the number of clusters, following figure compares the error value obtained with different values of n_clusters.

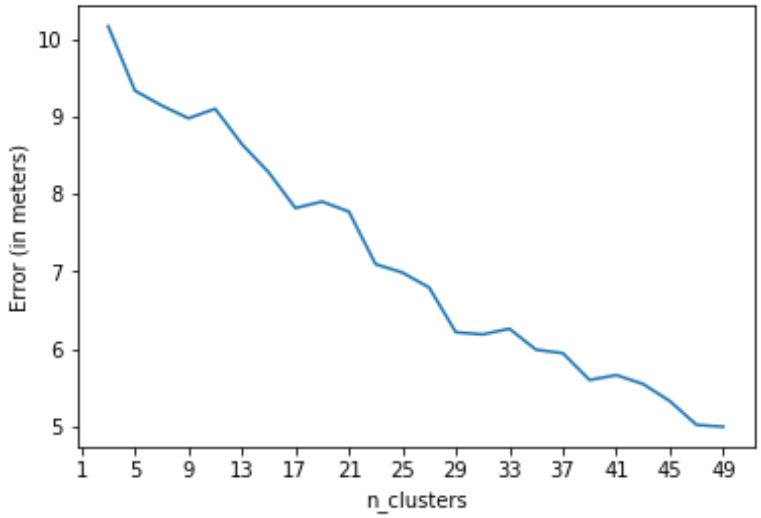


Fig. 4.7: Performance of KNN with clustering model with different number of clusters

On the other hand a high performance accuracy is obtained using the KNN classification model to predict the building number (99%) and floor number (88%) of the user.

The following table compares the results of regression models used to predict coordinates of the user.

Table 4.1: Algorithms with their positioning error

S. No.	**Model**	**Positioning error (in meters)**
1	Linear regression	53.95
2	Ridge regression	53.90
3	Lasso regression	50.80
4	KNN (K=1, default)	12.50
5	KNN (K=5, optimal)	11.04
6	KNN (K=49, worst)	12.95
7	KNN with K-means clustering (K=5, n_clusters=3)	10.16
8	KNN with K-means clustering (K=5, n_clusters=49)	4.99

From the above table it is evident that the proposed hybrid model i.e. KNN with clustering gives the best result thus by achieving the position accuracy of 5 meters in contrast to the 11 meters accuracy achieved by the solo KNN model deployed with an optimal value of K. While the KNN model improves the performance by 90% when the value of k is optimised from k=49 to k=5, the KNN and k-means hybrid algorithm further enhances the localization accuracy.

4.4.2 Comparing the accuracy of each model using CDF plots

In this dissertation we use the CDF plots for performance evaluation of our machine learning models. The empirical CDF is the proportion of values less than or equal to X. CDF plots are useful for comparing the distribution of different sets of data.

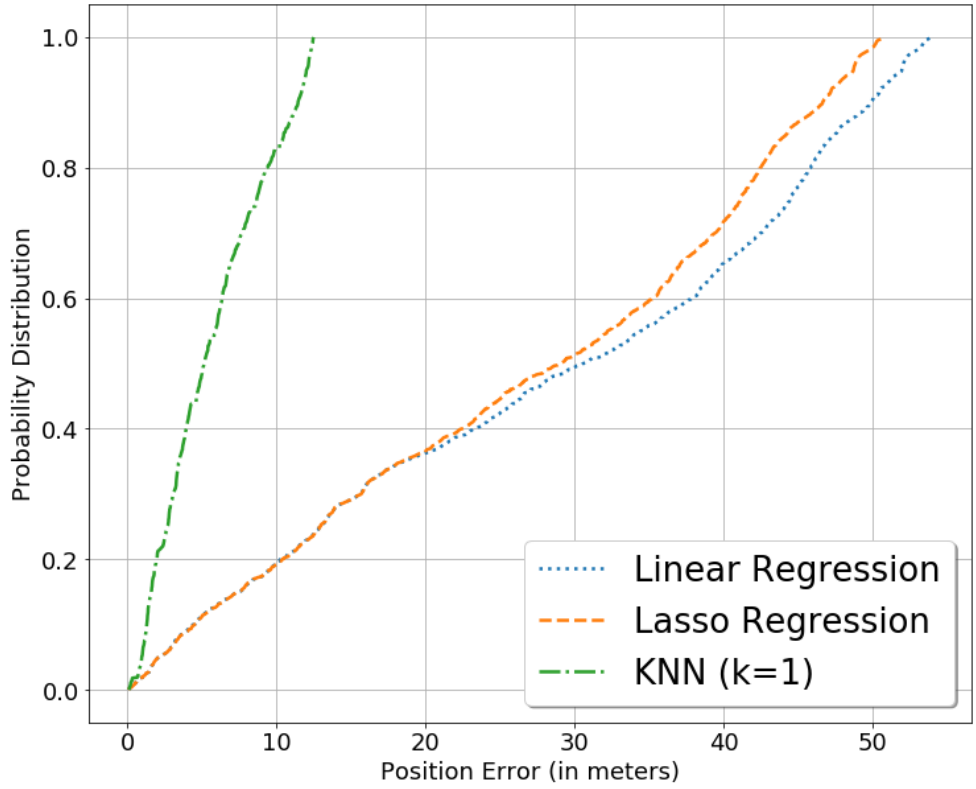


Fig. 4.8: Cumulative Density Plot representing the positional error vs the probability distribution of Linear Regression model, Lasso model and the default KNN model

It is evident from the above CDF plot that the traditional KNN with a default value of K performs much better than the linear regression as well as the lasso models. The poor performance of linear models can be linked to overfitting of training data.

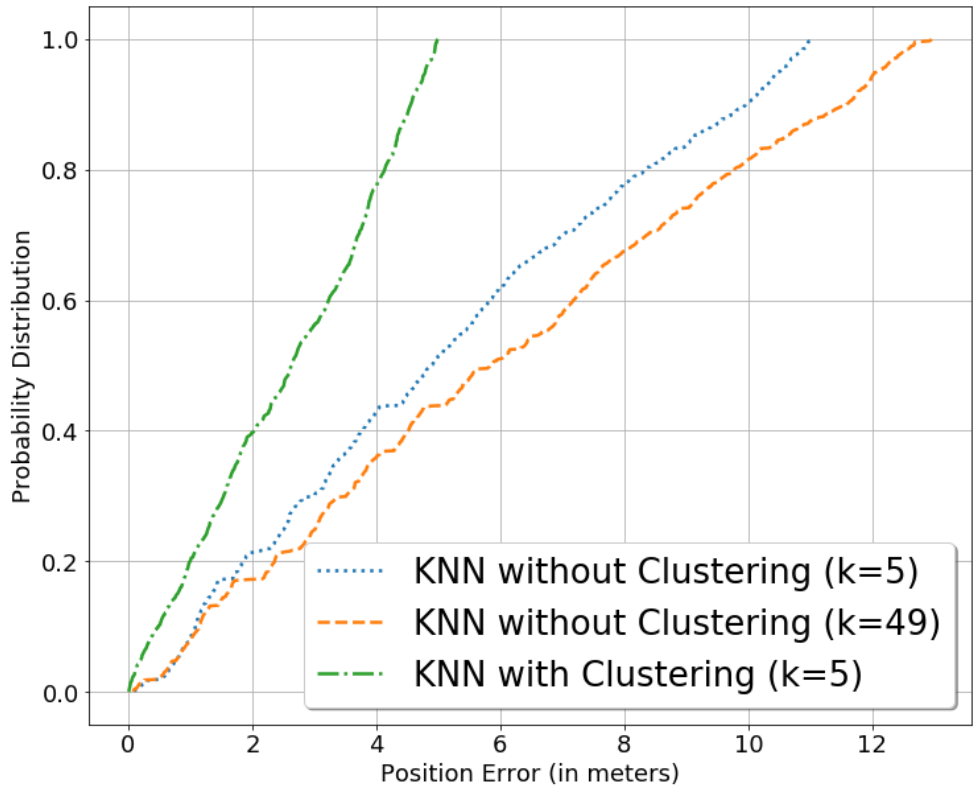


Fig. 4.9: Cumulative Density Plot representing the positional error vs the probability distribution of the KNN model with and without clustering

It can be clearly seen from the above figure that our proposed KNN with clustering model outperforms the traditional KNN with optimal value of K with a significant margin.

5 Summary

While empirical results and performance studies of positioning systems based on location fingerprinting have been presented in the literature, this dissertation has proposed a hybrid model for better performance on the Wi-Fi fingerprinting dataset.

Appropriate machine learning models used in this study were described in Chapter 3 after which the location fingerprint based on the received signal strength was investigated extensively in the next chapter. A systematic analysis was done to analyse the location fingerprint and discover its unique properties. We found that the RSS is random, with primarily a left-skewed distribution. In some cases, it is possible to approximate the RSS as being normally distributed. We used these assumptions to develop a model that can be used with analytical expressions for correctly estimating the position coordinates of a mobile device with a machine learning algorithm. Accuracy and precision are the two major performance metrics in this study. The following sections conclude the results obtained in Chapter 4 followed by a brief about the future scope of our study.

5.1 Conclusion

The performances of indoor positioning systems are mainly dependent on the precision of their positioning techniques and the algorithms used. As Wireless LAN (WLAN) costs less and is easy to access when compared with other indoor positioning techniques (IR, Ultrasonic, RFID, etc.), using WLAN for indoor positioning has become one hot topic of research.

In this dissertation, we briefly discussed some of the main methods of Wi-Fi fingerprinting based indoor positioning. We further pointed out that the typical KNN model could be improved if some selective pre-processing could be done. To validate this, we proposed a hybrid algorithm of using KNN with clustering, which utilizes clustering to filter out some of the neighbours. And finally, with the experimental results, we showed that KNN with clustering does outperform all other models.

The empirical results demonstrated that as the fingerprint data gradually increases it directly affects the positioning efficiency and the processing time, resulting in reducing positioning accuracy, so the optimization of fingerprint data is particularly necessary. Clustering evaluates the fingerprint data based on their degree of similarity and classifies them by it, resulting in optimizing the fingerprint data. After clustering, the fingerprint database can be divided into several small parts, improving the efficiency of positioning and reducing the system consumption.

Proposed positioning solution using K-means clustering in combination with the KNN method outperformed the traditional KNN. The K-means clustering reduces the required reference data and decreases the positioning average distance errors.

5.2 Further research

Based on this work, we will make efforts to establish a reliable enough indoor positioning system and finally put it in use for our further study or everyday life. Some of the possible paths for further research are explained below.

5.2.1 Providing location estimations with some confidence measure

The advantage of using the confidence measure can be evaluated to assess whether it produces valid predictions, and if there is any improvement on the positioning result. A confidence level can be used to represent the uncertainty of the positioning prediction. The concept of confidence machine is that a prediction made by any learning algorithm should be governed by a confidence parameter measuring the belief of the algorithm on this prediction. The confidence learning algorithm called conformal prediction (CP), produces a set of predictions, given a new sample, a training database and a confidence level. For instance, given a 95% confidence level, a training database with three training examples and a new sample s, the set of predictions that CP produces can be interpreted as 'I am 95% confident that "s" belongs to this prediction set. However, there is 5% chance that I may be wrong'.

5.2.2 Extending the learning algorithm to an online setting

For indoor positioning mechanisms, it is also possible to incorporate collaboration in order to improve system performance, especially for fingerprinting-based approaches. Active Campus [47] is an early system integrating user feedback. It allows users to update the training data incrementally for future use. When the system location is incorrect, users can click on the correct location and suggest new positions. The participation of end-users can actually assist in the construction of a positioning system incrementally from scratch. It is also evident that the user feedback-based positioning system adapts quite well when surroundings change.

Two more interesting research paths are possible. First, the unified performance evaluation methodology for all indoor positioning system based on location fingerprint is needed to allow a fair comparison among variety of emerging indoor positioning systems. So far, the accuracy and precision are the only performance metrics used for comparison. Second, a study of the indoor positioning system on with multi-floor and three-dimensional coordinates is not available. The impact of multiple floors is not yet known.

5.3 Self-assessment

While working on my dissertation, I got to learn a lot about new things. Researching the indoor positioning problem and how to solve it with the Wi-Fi fingerprinting technique using machine learning gave me an opportunity to dive deeper into the machine learning world and it's applications. This project required me to perform data analysis, pre-processing, data visualisation and finally fitting the models to predict the labels, which is an ideal job for a machine learning researcher.

I really enjoyed working with the most widely used machine learning algorithms, especially the part of implementing them from scratch at the start of the project. I am really grateful to Dr. Khuong An Nguyen, my project supervisor who gave me the idea of driving my own models and verifying them using the sample dummy sets. This helped me to get a thorough understanding of the models used in this work and later made the task of applying them to the competition dataset much easier.

This project helped me identify data pre-processing as one of the key stage in applying machine learning algorithms. A much better model was derived after the data pre-processing step resulting in very accurate results. I got to use various data visualisation techniques for comparing the results and getting insights from the data which is a useful tool for data analysis.

Finally, along with parting me the technical knowledge this project helped me hone my soft-skills of organising the tasks in an efficient and productive way. Managing the source code on GIT (version control system), taking meeting minutes, explaining and concluding the empirical results helped me improve my organising and writing skills.

References

[1] Krishnamurthy, Prashant. "Position location in mobile environments." *NSF Workshop on Context-Aware Mobile Database Management (CAMM)*. 2002.

[2] Rappaport, Theodore S., Jeffrey H. Reed, and Brian D. Woerner. "Position location using wireless communications on highways of the future." *IEEE communications Magazine*34.10 (1996): 33-41.

[3] Pahlavan, Kaveh, Xinrong Li, and Juha-Pekka Makela. "Indoor geolocation science and technology." *IEEE Communications Magazine* 40.2 (2002): 112-118.

[4] Hightower, Jeffrey, and Gaetano Borriello. "Location systems for ubiquitous computing." *Computer* 8 (2001): 57-66.

[5] Ladd, Andrew M., et al. "Robotics-based location sensing using wireless ethernet." *Wireless Networks* 11.1-2 (2005): 189-204.

[6] Weiser, Mark. "Some computer science issues in ubiquitous computing." *Communications of the ACM* 36.7 (1993): 75-84.

[7] Ramani, Ishwar, Rajiv Bharadwaja, and P. Venkat Rangan. "Location tracking for media appliances in wireless home networks." *2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698)*. Vol. 2. IEEE, 2003.

[8] Djuknic, Goran M., and Robert E. Richton. "Geolocation and assisted GPS." *Computer* 34.2 (2001): 123-125.

[9] Bahl, Paramvir, et al. "RADAR: An in-building RF-based user location and tracking system." (2000).

[10] Tauber, J. A. "Location systems for pervasive computing." *Area Exam Report, Massachusetts Institute of Technology*(2002).

[11] Meyer, Michael J., et al. "Wireless enhanced 9-1-1 service—making it a reality." *Bell Labs Technical Journal* 1.2 (1996): 188-202.

[12] Brunato, Mauro, and Roberto Battiti. "Statistical learning theory for location fingerprinting in wireless LANs." *Computer Networks* 47.6 (2005): 825-845.

[13] Roos, Teemu, et al. "A probabilistic approach to WLAN user location estimation." *International Journal of Wireless Information Networks* 9.3 (2002): 155-164.

[14] Hightower, Jeffrey, Barry Brumitt, and Gaetano Borriello. "The location stack: A layered model for location in ubiquitous computing." *Proceedings Fourth IEEE Workshop on Mobile Computing Systems and Applications*. IEEE, 2002.

[15] Want, Roy, et al. "The active badge location system." *ACM Transactions on Information Systems (TOIS)* 10.1 (1992): 91-102.

[16] Ward, Andy, Alan Jones, and Andy Hopper. "A new location technique for the active office." *IEEE Personal communications* 4.5 (1997): 42-47.

[17] Priyantha, Nissanka B., Anit Chakraborty, and Hari Balakrishnan. "The cricket location-support system." *Proceedings of the 6th annual international conference on Mobile computing and networking*. ACM, 2000.

[18] Ni, Lionel M., et al. "LANDMARC: indoor location sensing using active RFID." *Proceedings of the First IEEE International Conference on Pervasive Computing and Communications, 2003.(PerCom 2003).*. IEEE, 2003.

[19] Lacage, Mathieu, Mohammad Hossein Manshaei, and Thierry Turletti. "IEEE 802.11 rate adaptation: a practical approach." *Proceedings of the 7th ACM international symposium on Modeling, analysis and simulation of wireless and mobile systems*. ACM, 2004.

[20] Fukuju, Yasuhiro, et al. "DOLPHIN: An Autonomous Indoor Positioning System in Ubiquitous Computing Environment." *WSTFES* 3 (2003): 53.

[21] Harle, Robert. "A survey of indoor inertial positioning systems for pedestrians." *IEEE Communications Surveys & Tutorials*15.3 (2013): 1281-1293.

[22] Klipp, Konstantin, et al. "Low cost high precision indoor localization system fusing intertial and magnetic field sensor data with radio beacons." *Second Annual Microsoft Indoor Localization Competition* (2015).

[23] Pahlavan, Kaveh, and Prashant Krishnamurthy. *Principles of wireless networks: A unified approach*. Prentice Hall PTR, 2011.

[24] Liu, Hongbo, et al. "Push the limit of WiFi based localization for smartphones." *Proceedings of the 18th annual international conference on Mobile computing and networking*. ACM, 2012.

[25] Wang, Zhi, et al. "AIDLOC: AN ACCURATE ACOUSTIC INDOOR LOCALIZATION SYSTEM."

[26] Ganick, Aaron, and Daniel Ryan. "Method and system for modulating a light source in a light based positioning system using a DC bias." U.S. Patent No. 8,334,901. 18 Dec. 2012.

[27] Kumar, Navina, et al. "Visible light communication systems conception and vidas." *IETE Technical Review* 25.6 (2008): 359-367.

[28] Brena, Ramon F., et al. "Evolution of indoor positioning technologies: A survey." *Journal of Sensors* 2017 (2017).

[29] Liu, Hui, et al. "Survey of wireless indoor positioning techniques and systems." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 37.6 (2007): 1067-1080.

[30] Gezici, Sinan, et al. "Localization via ultra-wideband radios: a look at positioning aspects for future sensor networks." *IEEE signal processing magazine* 22.4 (2005): 70-84.

[31] Mautz, Rainer. "Overview of current indoor positioning systems." *Geodezija ir kartografija* 35.1 (2009): 18-22.

[32] Bahl, Paramvir, et al. "RADAR: An in-building RF-based user location and tracking system." (2000).

[33] Saha, Siddhartha, et al. "Location determination of a mobile device using IEEE 802.11 b access point signals." *2003 IEEE Wireless Communications and Networking, 2003. WCNC 2003.*. Vol. 3. IEEE, 2003.

[34] Youssef, Moustafa A., Ashok Agrawala, and A. Udaya Shankar. "WLAN location determination via clustering and probability distributions." *Proceedings of the First IEEE International Conference on Pervasive Computing and Communications, 2003.(PerCom 2003).*. IEEE, 2003.

[35] Bahl, Paramvir, Venkata N. Padmanabhan, and Anand Balachandran. "A software system for locating mobile users: Design, evaluation, and lessons." *Microsoft Research, MSR-TR-2000-12* (2000).

[36] Van Nee, Richard, et al. "New high-rate wireless LAN standards." *IEEE Communications Magazine* 37.12 (1999): 82-88.

[37] Kaemarungsi, Kamol, and Prashant Krishnamurthy. "Modeling of indoor positioning systems based on location fingerprinting." *Ieee Infocom 2004*. Vol. 2. IEEE, 2004.

[38] Harrell Jr, Frank E., et al. "Regression modelling strategies for improved prognostic prediction." *Statistics in medicine* 3.2 (1984): 143-152.

[39] Breiman, Leo, and Jerome H. Friedman. "Predicting multivariate responses in multiple linear regression." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 59.1 (1997): 3-54.

[40] Schmidt, Mark, Glenn Fung, and Rmer Rosales. "Fast optimization methods for l1 regularization: A comparative study and two new approaches." *European Conference on Machine Learning*. Springer, Berlin, Heidelberg, 2007.

[41] Tibshirani, Robert. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996): 267-288.

[42] Hastie, Trevor, and Robert Tibshirani. "Discriminant adaptive nearest neighbor classification and regression." *Advances in Neural Information Processing Systems*. 1996.

[43] Wagstaff, Kiri, et al. "Constrained k-means clustering with background knowledge." *Icml*. Vol. 1. 2001.

[44] Berkvens, Rafael, Maarten Weyn, and Herbert Peremans. "Position error and entropy of probabilistic Wi-Fi fingerprinting in the UJIIndoorLoc dataset." *2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE, 2016.

[45] Ma, Jun, et al. "Cluster filtered KNN: A WLAN-based indoor positioning scheme." *2008 International Symposium on a World of Wireless, Mobile and Multimedia Networks*. IEEE, 2008.

[46] Torres-Sospedra, Joaquín, et al. "UJIIndoorLoc: A new multi-building and multi-floor database for WLAN fingerprint-based indoor localization problems." *2014 international conference on indoor positioning and indoor navigation (IPIN)*. IEEE, 2014.

[47] Bhasker, Ezekiel S., Steven W. Brown, and William G. Griswold. "Employing user feedback for fast, accurate, low-maintenance geolocationing." *Second IEEE Annual Conference on Pervasive Computing and Communications, 2004. Proceedings of the*. IEEE, 2004.

[48] Sklar, Bernard. "Rayleigh fading channels in mobile digital communication systems. I. Characterization." *IEEE Communications magazine* 35.7 (1997): 90-100.

[49] Smailagic, Asim, Jason Small, and Daniel P. Siewiorek. "Determining user location for context aware computing through the use of a wireless LAN infrastructure." *Institute for Complex Engineered Systems Carnegie Mellon University, Pittsburgh, PA* 15213 (2000).

A Hardware setup and how to run it?

The analysis is done using python as the main programming language. Various python libraries are used for the visualisation and model fitting purposes including- scikit-learn, matplotlib, pandas, numpy.

Jupyter Notebook is used for the analysis purpose, which can simply be run on a machine with python-3 installed on it and running on a conda environment.

- **Recommended System Requirements**:
  - Processors: Intel® Core™ i5 processor 4300M at 2.60 GHz or 2.59 GHz (1 socket, 2 cores, 2 threads per core), 8 GB of DRAMIntel® Xeon® processor E5-2698 v3 at 2.30 GHz.
  - Disk space: 2 to 3 GB.
  - Operating systems: Windows® 10, macOS, and Linux.
  - Python version: 3.6.X.
- **Minimum System Requirements**:
  - Processors: Intel Atom® processor or Intel® Core™ i3 processor.
  - Disk space: 1 GB.
  - Operating systems: Windows 7 or later, macOS, and Linux.
  - Python versions: 2.7.X, 3.6.X.

B Professional issues

In this this section we will mention some of the professional issues related to this project on Wi-Fi fingerprinting. Privacy of the user is the main concern for location based services. Location-based data is actually sensitive information that must not be dealt with lightly. In many cases the functionality could be exploited and the providing system could be abused.

Location information should be made available only to those with authorized access. This issue represents the privacy concern of mobile users who do not want to reveal their location or be tracked. It is closely related to how the system determines the location information and the type of application. A system similar to the GPS where each GPS device derives its own position from the GPS satellites can completely secure the user location information. On the other hand, a location tracking such as the E-911 system [11] with the main purpose to capture the user location can be abused by unauthorized groups if there is no security protection in place.

Personalization is one of the key features of intelligent, context-aware, adaptive indoor positioning systems. However, this requires the storage of personal preferences, activity history, current location and previous movements. The threats associated with the violation of location privacy can dramatically limit the development, adoption and growth of indoor positioning applications. It require the user to disclose their location to enable personalization.

Service providers can potentially store, use (or misuse, reuse), and sell location data. Such potential threats can discourage users. Unrestricted access to information about an individual's location could potentially lead to harmful encounters. In addition, an individual's location history can potentially disclose activities, preferences, health, background and history and other (even more) private aspects of life. In particular, if the locations are accompanied by temporal information, the trajectory of movement, then more can be revealed.

Exposure of user information can lead to different damage for both users and service providers. Therefore, it is important to manage private data securely; not only applying the existing methods of encryption in communications and databases, but also providing a solution for hiding users' information from different kind of risks. The location system should have a security protocol embedded within the system to protect the location information. Unfortunately, the security of the system is limited by the location sensing technique. For instance, a positioning system that reuses the communication signals for the purpose of location detection cannot completely secure the mobile station's privacy because of its active nature.

On the other hand a strict and more private system would mean that there is a very little scope for getting the user-feedback or using the reinforcement learning technique to improve our algorithms. This would ultimately decrease the data collection and result in a slow progression of the system.

There is a need of a system which has a perfect balance between the above mentioned issues which calls for further research on the currently available solutions and replacement of those posing a serious potential threat to user privacy.

C Code

All the working code produced and presented for this analysis is provided together with this document. Otherwise, the code can be found on git at the following link.

https://github.com/handabaldeep/Wifi-fingerprinting