Universidade do Minho
Escola de Engenharia
Departamento de Informática

David Kramer

Crowd Sensing for Urban Security in Smart Cities

January 2020

Universidade do Minho
Escola de Engenharia
Departamento de Informática

David Kramer

Crowd Sensing for Urban Security in Smart Cities

Master dissertation
Master Degree in Computer Science

Dissertation supervised by
Prof. Cesar Analide
Prof. Bruno Fernandes

January 2020

ABSTRACT

Bringing intelligence to our everyday environments is a growing reality and therefor ewe should take advantage of the technology available to improve several areas of our daily life. For example, current technology allows the conception of smart scanners to passively detect devices such as smartphones or smartwatches. Developing CrowdControl algorithms and making use of these Crowd Sensors will enable the gathering of important data which will contribute for the continuous improvement of Urban Security. Considering the Vulnerable Road Users' (VRUs) problem, the goal is to sense the density of people at certain points of interest for VRUs through the use of a CrowdSensor, allowing, for example, the better location of crosswalks or the identification of crowded spots, enhancing the security of all VRUs. More specifically, this dissertation aims to use smart scanners to passively detect smartphones and smartwatches throughprobe requests emitted by such devices. Then, besides the use of Crowd Detection and Control algorithms, the goal is to get the distance of a person to the smart scanners using the RSSI (Received Signal Strength Indicator) and conceive and develop Machine Learning models to forecast the density of the sensed areas.

Keywords: Ambient Intelligence, Artificial Intelligence, Smart Cities, Internet of Things, Machine Learning, Crowd Sensing, Wi-Fi Probe Requests, Vulnerable Road Users

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AI Artificial Intelligence. 1, 7, 26
AmI Ambient Intelligence. 1, 2, 6–9, 12, 33, 34, 39
ANNs Artificial Neural Networks. 1, 30
ARIMA Autoregressive integrated moving average. 1, 30

CNN Convolutional Neural Network. 1, 30

ETS Exponential smoothing state space model. 1, 21, 30

GPS Global Positioning System. 1, 11

IoT Internet of Things. 1, 5, 6, 23

LSTM Long Short-Term Memory. 1, 30

MCS Mobile Crowdsensing. 1, 23, 24
MEI Mestrado em Engenharia Informática. 1
ML Machine Learning. 1, 3, 4, 8, 19, 24, 26–30, 34, 39

OUI Organizationally Unique Identifier. 1

PCA Principal Component Analysis. 1, 29

RFID Radio-frequency identification. 1, 7, 23
RNN Recurrent Neural Network. 1, 30
RPD Pre-Dissertation Report. 1
RSSI Received Signal Strength Indicator. 1, 3, 8, 12, 15, 17, 18, 22, 24, 31, 34, 35

SOM Self Organizing Map. iii, 1, 19, 20, 30

U-matrix Unified distance matrix. 1, 20
UM Universidade do Minho. 1

VRUs Vulnerable Road Users. 1, 2, 33

WPAN Wireless Personal Area Networks. 1, 11

1

INTRODUCTION

In this section, the context and motivation of the problem will be discussed, along with the definition of the project's main objetives and research methodology. Finally, it will give an overview of this paper's structure.

1.1 CONTEXT & MOTIVATION

Bringing intelligence to our everyday environments is a growing reality and therefore we should take advantage of the technology available to improve several areas of our daily life. For example, one potential area to apply this study on is Urban Security, in which the Vulnerable Road Users (VRUs) problem stands out because of its importance. The term "Vulnerable Road Users" refers to non-motorised road users, such as pedestrians , cyclists as well as motor-cyclists and persons with disabilities or reduced mobility and orientation (E. P. C. European Union). According to the World Health Organization, the number of road traffic deaths is increasing every year, which shows the relevance of this problem. The growing reality of technology, namely Ambient Intelligence (AmI) and Smart Cities, respectively defined as *"a digital environment that proactively, but sensibly, supports people in their daily lives"* Cook et al. (2009) and as a city that has the *"ability to reason upon the knowledge acquired through data gathered by sensorization, with focus on improving the quality of life at urban centres, considering sustainability and safety principles"* (Fernandes et al., 2018), enables the development of different solutions to improve areas like urban Security, marketing or energy management. For example, current technology allows the conception of smart scanners to passively detect devices such as smartphones or smartwatches through the sense of Wi-Fi probe requests and Bluetooth signals. Developing Crowd Control algorithms and making use of these Crowd Sensors will enable the gathering of important data (like crowd density data) which will contribute for the continuous improvement of Urban Security, for example (Fernandes et al., 2018; Cook et al., 2009; Jung and Muñoz, 2018). In terms of motivation, the following vision by Weiser (1991) represents the author's motivation for this project: *"The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it"*. The main goal and motivation of this project is to develop a system that has the ability to improve areas like Urban Security whilst the user does not have an active role in the gathering of data, enabling the user to live his daily life passively and having its Urban Security improved significantly. The fact that this project also involves the use and theoretical knowledge of emerging technologies, like ML, and a possible contribution for the affirmation of the concept of Smart Cities, also plays a big role in the development of this project.

1.2 MAIN OBJECTIVES

Considering the VRUs problem, the goal is to sense the density of people at certain points of interest for VRUs through the use of a Crowd Sensor, allowing, for example, the better location of crosswalks or the identification of crowded spots, enhancing globally the Urban Security. More specifically, this dissertation aims to firstly choose an adequate point of interest in which smart scanners will be used to passively detect smartphones and smartwatches through probe requests emitted by such devices, estimating the crowd density of that area. Then, besides the use of Crowd Detection and Control algorithms, the goal is to get the distance of a person to the smart scanners using the Received Signal Strength Indicator (RSSI) and conceive and develop Machine Learning models to forecast the density of the sensed areas.

1.3 RESEARCH HYPOTHESIS

On one hand, it is expected to prove that the use of smart scanners and Wi-Fi probe requests are a viable and accurate method to estimate and forecast crowd densities without any active role of the user being sensed. On the other hand, it is expected that this study can improve the quality of life in several areas. Depending on where the sensors are going to be deployed, the sensed data can be useful for a range of applications: traffic forecasting, better crowd safety measures or evaluating whether a place is suitable for a shop or not. Finally, developing a system capable of enhancing one's life without actively taking any part in it is also a big objective of this dissertation.

1.4 PAPER STRUCTURE

This paper will contain the following main sections: Introduction, State of Art, The Problem and Its Challenges and all the development done so far.

Firstly, in the introductory section it will establish an overall view on this thesis theme context, motivation and main objectives.

In the State of Art section, the goal is to describe minutely the different kind of approaches on crowdsensing, specially the approaches that also involve the detection of Wi-Fi or Bluetooth probe requests and then reason upon its main advantages and disadvantages. Besides the description of the different solutions to help enhancing the numerous areas of a Smart City, this section will also contain information about the current state of Machine Learning (ML), along with its different paradigms.

In the next section, the problem and its challenges are going to be explained from different perspectives so that it's possible to see what kind of issues might arise in future work.

In the development section, it will be presented the work methodology, the work done so far, technologies used and future work. More specifically,in the work done so far part a detailed overview of the developed code will be given, along with information about the sensing device used and the software chosen to do so. Additionally a quick data analysis will be done using a dataset expected to be similar or equivalent to the one being produced by the proposed solution of this dissertation.

Finally, the last section will present the conclusions of the work done so far along with the the steps to be taken in the next months.

2

STATE OF THE ART

In this section, it will be explained how Smart Cities and Ambient Intelligence are being broadly used and studied nowadays, and more specifically, how different authors approach crowdsensing and how they make the sensed data useful for different case studies.

2.1 SMART CITIES

In the cities' history, the importance of technology to improve quality of life has been always recognized and explored. Over the humankind's history, there are several examples of how technology can be useful for cities. Indeed, 3,000 years ago Persian Enginneers dug a Qanat, a long tunnel connecting a well to its outflow many miles away. This ancient example of open and forward thinking, made possible the water supply for a city of one million inhabitants until very recently (Harrison et al., 2010).

In this modern times, with all the technology available we should strive for making better use of all the public resources available in order to improve numerous areas of a city and ultimately increase the quality of life of each citizen along with the reducement of operational costs of the public administrations (Zanella et al., 2014). Along side with Smart Cities stands Internet of Things (IoT) (Fernandes et al., 2018). IoT is now related with technologies such as sensors, actuators, GPS devices, and mobile devices (Xu et al., 2014). The use of these resources can improve the quality of areas such as agriculture, urban security, environmental monitoring, security surveillance, and others (Xu et al., 2014; Fernandes et al., 2018; Zanella et al., 2014). In fact, the availability of different types of data collect by IoT sensors, may be used to enhance the awareness of people about the status of their city or to improve other urban issues (Zanella et al., 2014). For example, urban IoT could enable the route planning in advance to reach the office or to discover the time of day in which one could do a shopping trip to the city centre without encountering traffic. The sensing capabilities and GPS installed on modern vehicles along with the sensing of different conditions in a given road, would allow these type of situations to happen (Zanella et al., 2014).

One important aspect to mention is that despite the fact that these kind of technologies are still in a early stage, there are real life examples of already implemented systems. Recently, the widely known BMW, developed an intelligent informatics system (iDrive system) that provides intelligent driving directions based on environment data (vehicle location and road condition) provided by various sensors and tags (Qin et al., 2013).

Finally, Smart Cities and IoT have proven to be a great asset to improve different areas of a city and a citizen's daily life. However, with great power comes great responsibility. Given the fact that, for example, every IoT technology will rely largely on the collection of personal and private information, protecting this data is a crucial and important for the evolution of Smart Cities. In fact, studies have revealed that the number of attacks on IoT entities, when compared to the traditional ones, is appearing to be much higher (Roman et al., 2011; Li, 2013; Ting and Ip, 2013). Therefore, IoT standardization, such as the definition of privacy and legal interpretation, which are still not clearly defined, should be worked upon to prevent privacy issues (Xu et al., 2014).

2.2 AMBIENT INTELLIGENCE

Additionally, AmI should be given importance for the promotion of Smart Cities (Fernandes et al., 2018). AmI is a digital environment that brings intelligence and improvement to our daily lives by continously acting and reasoning upon the sensed data (Cook et al., 2009). In fact, a Smart City should embed AmI by creating several distinct sensorization levels in cities, as shown in picture 1. At a first level, there are APIs that offer relevant information such as road conditions and the weather. On a second level people can wear devices that makes them a citizen sensor, contributing for the extraction of relevant data. The third and final level comprises the city in itself and the methods for extracting actionable data from the environment in where one lives and stands (Fernandes et al., 2018).

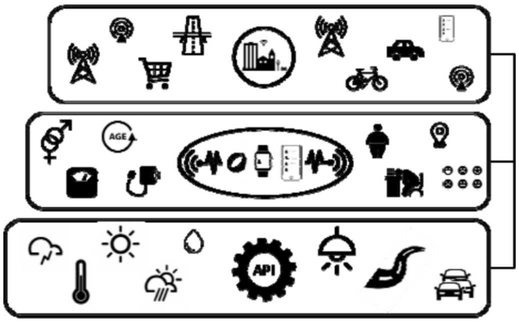


Figure 1: The scale levels for data collection and Ambient Intelligence in Fernandes et al. (2018).

When it comes to the AmI evolution, the evolution of technology played, naturally, a very important role. Initially computers were very expensive and difficult to understand and a single computer would typically be used by numerous individuals. This paradigm changed in the 80s with the PC revolution, allowing each individual to possess its own computer. Nowadays, as industry progressed and costs dropped, one individual often has access to more than one computer - naturally more complex and capable than the previous ones. Moreover, in current times the access is not limited to only just computers. This means that since the miniaturization of microprocessors, computing power is embedded in our daily lives. Whether present in mobile phones or cars, these devices are often something that we take with to travel outside our homes (like mobile phones or smartwatches) or to help us finding the best route to our destination (like cars and their GPS Navigation). All this evolution, namely in faster computation with reduced power along with an increased availability, made the realization of AmI possible Cook et al. (2009).

AmI algorithms normally follow the same logic. Firstly, the objective is to gather data by sensing the environment's and users' state with the use of sensors. Following up, the use of AI techniques allows the reasoning upon the sensed data. Finally, acting has the objective to make AmI algorithm reach its goal, by executing actions that affect the end users Cook et al. (2009). Shortly, AmI has three phases: **sensing**, **reasoning** and **acting**.

Sensing

Sensing is where AmI starts. In order to have some sort of effect on a individual's daily life, it is necessary to firstly collect data about this individual (accordingly to the purpose). This way, sensing is the process of perceiving the environment by collecting data effectively so that the reasoning and acting algorithms have pratical use (Cook et al., 2009). Effectively means not only to sense with the most accuracy possible, but also to have the adequate data processing.

Sensors can be used for different purposes and in different ways. Some sensors have been designed for the detection of chemicals and humidity sensing (G.Delapierre et al., 1983) or to determine readings for light, radiation, temperature, sound, strain, pressure, position, velocity, and direction, and physiological sensing to support health monitoring (Ermes et al., 2008; Stanford, 2004). Others were built to detect Bluetooth-enabled mobile phones carried by the participants in a festival (Larsen et al., 2013). Alternatively, a way of detecting individuals, can also be by the use of wearable devices, such as Radio-frequency identification tags that along with RFID readers enable the monitorization of the tagged objects. Also, the use of video-based techniques has been used for sensing purposes (Yuan et al., 2011).

Reasoning

Reasoning must operate with the data gathered by the sensing. In order to make this data useful, reasoning algorithms must be responsive, adaptive and beneficial to the problem that is being addressed. Therefore, these algorithms include different types of reasoning (Cook et al., 2009). They can vary from Modeling or Decision Making, to, in a more specific way, algorithms that enable crowd density estimation.

In terms of Modeling, the main goal is to build a model that is adaptable to a certain problem, by detecting anomalies and changes in patterns. As for Decision Making, the goal is to have automated decisions. For example, according to Mozer (2004), one of the few fully-implemented AmI systems, has the goal to determine ideal settings for lights and fans in a home, by using neural network and a reinforcement learner. Typically most of these algorithms use Machine Learning techniques, as they need to be responsive to changes and perfom prediction tasks.

Concerning crowd density estimation, there are several approaches that one can take. In the case of Schauer et al. (2014), in order to reduce the number of false-positives when sensing the crowd, a hybrid approach is used, considering both the RSSI value and the time when a MAC address was captured. By using two sensors, and comparing the time delay of a certain MAC address between the two sensors along with a certain threshold defined for RSSI, the authors prove that the possibility to reduce the false-positive rate is increased.

Acting

The last typical phase of a AmI system, is Acting. It allows the connection between the real world and all the work done by the previous phases through the execution of actions that affect the system users. It can be done by robots or simply by notifications or interactions (Cook et al., 2009; Ramos et al., 2008).

Mobile and Pervasive Computing

One main goal of AmI is to make pervasive computing, ubiquitous computing and context-aware computing a reality. Indeed, shortening the gap between humans and computing devices, would allow the wide spread of this technologies through a area or a group of people, making possible the development of devices without explicit operator control Cook et al. (2009); C.Augusto et al. (2010). In the context of this dissertation, it is crucial to reduce the human-computer interaction, so that the system can infer situations without the active role of any person.

Finally, AmI is a paradigm that can bring multiple benefits to our daily lives and in different ways while still being adaptive, as shown above. Every phase has its own purpose and are mutually dependent. Nevertheless, similarly to Smart Cities, there are privacy and security challenges that must be acknowledged. On one hand, the fact that every AmI system requires some sort of sensing on a user's environment or daily life, even with potential benefits, brings up privacy issues. Moreover, in 2003 a survey showed that privacy protection was more important to the participants than any potential benefits provided AmI technologies. Nonetheless, there is also a city whose evolution was enhanced by the open acceptance to loss of privacy (O'Connell, 2005). On the other hand, issues like the AmI systems performing the wrong actions and forcing humans to extra work - having the opposite intended effect - are an aspect that makes the deployment of these systems a process that requires caution.

2.3 WI-FI PROBING

These days, the majority of smart devices are equipped with Wi-Fi communication interfaces. This type of interface is normally used for enabling internet access, and has been widely adopted by any kind of smart device. In recent years, smart devices usage has suffered a huge increase (Zhou, 2017/09; E.Longo et al., 2018). For example, in 2016, the shipment of China's smartphones represented 95.7% of China's total mobile phones shipment in the same period (Zhou, 2017/09). Other valuable facts are that the number of cellphones is now bigger than the actual number of people in the world, according to *UN's International Telecommunications Union (ITU)*, the *World Bank*, and the UN (United Nations) (M.Murphy, 2019), or that the number of smartphones users worldwide is increasing every year, as show in picture 2. Therefore, Wi-Fi has been widely used and its usage is increasing at a fast pace.

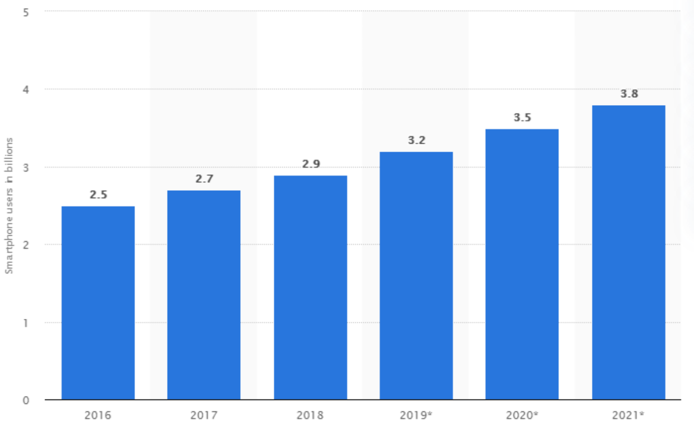


Figure 2: Number of smartphone users worldwide from 2016 to 2021 in Holst (2019).

Smart devices that have a Wi-Fi interface periodically send Wi-Fi probe requests in order to discover which wireless networks are available for connection (L.Oliveira et al., 2019). This mechanism is defined by the IEEE 802.11 standard as an active mechanism or an active scan (Zhou, 2017/09; Freudiger, 2015). These probe requests are constantly being sent whenever the smart device's Wi-Fi is enabled, sending information containing for example the device's MAC address. It should be noted that the rate at which a device sends this packets depends on the OS used on it (Freudiger, 2015). To sum up, this communication process that happens every time the device's Wi-Fi is enabled, allows the sniffing of the probe requests, allowing the detection of devices nearby a certain sensor (Zhou, 2017/09). This opens the possibility to use smart devices' probe requests as a way to get insights into numerous areas like crowd density (Schauer et al., 2014), shopping habits (Barbera et al., 2013) or traffic forecasting (Ares et al., 2016).

Over the past few years, the MAC address, which is sent along a probe request, has suffered changes due to privacy issues. At first, the MAC address of each smart device was unique and immutable, which lead to a lot of contest regarding privacy dangers. Nowadays, the main strategy adopted by manufactures to fight this issue, is the randomization of the MAC addresses (Freudiger, 2015; L.Oliveira et al., 2019). There were other proposed approaches, which did not have continuity: Beresford and Stajano (2004) suggested changing MAC addresses over time, whereas Greenstein et al. (2008) proposed removing link-layer identifiers al-together. In L.Oliveira et al. (2019) the randomization of MAC addresses is defined as *a strategy that is intended to prevent potential observers from identifying which mobile devices are within reach of a sensor*. Moreover, details like device, manufacturer, and operating system version influence how the MAC address is or not randomized (L.Oliveira et al., 2019). Martin et al. (2017) identified that most devices, especially those with an Android operating system, do not implement randomization of MACs in any way. There are also studies that show that, the MAC randomization, in the case the iOS 8.1.3 randomization mechanism, can be defeated (Freudiger, 2015). Besides that, others have showed that is possible to estimate the number of mobile devices present at a certain place and time through a solution that is immune to Media Access Control (MAC) address randomization strategies (L.Oliveira et al., 2019).

To conclude, Wi-Fi is a growing technology that is increasingly more integrated in ubiquitous computing. Its usage in most of today's devices, opens new possibilities on how to implement crowdsensing. Even though MAC address randomization is feature that does protect user privacy, it is still an on-going process, as many devices still don't have it implemented, or because there are methods to get around it. Therefore, finding a way to implement crowdsensing without compromising one's privacy while still getting the most of the fact that smart devices are constantly sending probe requests, is one of the objectives of this dissertation.

2.4 CROWDSENSING

One big goal of this dissertation is to sense the crowd, gathering data either for calculating density of people at certain points of interest, or even forecast these values. Crowdsensing can be done in multiple different ways, and consequently, the objective of this section is to review different approaches to this problem. There are some key aspects which will be taken into account in order to have some comparison parameters:

1. Sensing
    - The kind of device used for the sensing;
    - Technologies involved;
    - The collected data structure;
    - The role of the user in the sensing (an inactive role or opportunistic and participatory sensing) (Sun et al., 2016; Petkovics et al., 2015; Ganti et al., 2011; Guo et al., 2015);
    - The scale on which the approach is being tested;
    - The storage method used;
    - What kind of privacy measures are used;
    - Pre-processing.
2. Reasoning
    - Data Modeling;
    - Machine Learning Models used.

2.4.1 *Sensing through Bluetooth*

The paper produced by Larsen et al. (2013) presents an approach to Crowd Sensing at a large-scale music festival with over 130,000 participants for obtaining spatio-temporal data. In order to do so, the authors use several Blutetooth scanners to discover Bluetooth-enabled mobile phones carried by the participants. In a second phase, they reveal what kind of algorithms used for analysis purposes.

Methodology

In the referred study, the authors' method to obtain data is to use **Nokia N900** smartphones with custom software for detecting Bluetooth-enabled devices in proximity, as crowd sensors. As the paper Larsen et al. (2013) states, this methods has some benefits and drawbacks. On one hand, the Bluetooth scanners, functioning as master devices, scan passively continuously for devices without any active participation on the user side. On the other hand, since Bluetooth is a short-range low-power protocol for implementing Wireless Personal Area Networks, the range in which devices can be discovered is very limited. Besides that, the discoverability of a device can depend on its operating system and additionally on the fact that normally it has to be set manually by the user.

Regarding the scanner itself, besides having the Bluetooth module in it, it offered 3G communication, data storage, battery power and Global Positioning System tracking. This allowed the gather of data in real time whilst having power in case of a event's power outage, or the tracking of the device in case the it got lost.

Data

In terms of data storage, the scanned data was stored locally using the SQLite database on the device and also uploaded to server, always depending on the network availability. In order to maintain availability and robustness of the system, the authors present a not so clear approach. Approximately there were two scans per minute, and if the devices did not upload data to the server during a certain amount of a time, the device would be rebooted either by issuing a command via Bluetooth or by manually turning them on and off. It is not clear if the command via Bluetooth was triggered automatically by the software itself. Additionally, having any kind of manual work in this case is not the best solution when the objective is to maintain availability and to implement an AmI system. Nevertheless, an periodical reboot would occur every 24 hours to minimize the effect of any device not working properly.

As for the data structure, the collected data is a time-series of events, where each entry is characterized by the time, scanner ID and the Bluetooth MAC address of a discovered device. The authors argue that the RSSI was not registered, since using that measure for distance calculation has an accuracy which can depend on the type of environment (Larsen et al., 2013).

Regarding privacy issues, even that according to Larsen et al. (2013) the Danish Data Inspectorate considered that that information didn't enable the linkage of the device to a person, the authors still made sure to hash the MAC addresses after extracting information about the vendor. Also, the human-readable identifiers on each device were not recorded for a faster scanning and for anonymity issues.

Data collection and Analysis

In order to provide sufficient coverage of all the relevant areas, the authors used 33 Bluetooth scanners placed in the vicinity of stages like shops, bee booths and mixing areas of the stages. This allowed the coverage of rich spots and also the placement in spots where a power source was available. The scanned data was uploaded in real time via a 3G network. Naturally, in events where high number of mobile phones are present problems with the mobile network tend to occur. So some sensors only uploaded data when they got their connection reestablished- normally in the early morning hours. With this being said, only 7 of 33 sensors were able to run without the need of being maintained one or more times during the 8 days of the festival. With this strategy, the authors were able to collect 1,204,725 observations during the 8 days of the festival. Taking into account the number of unique devices discovered (8,534), 6,5% of the overall population was actually observed. Therefore, 8,534 entries in the dataset does provide some window to discover patterns in participants mobility.

Micro Groups Modelling

Taking into account that the authors claim that given the fact that the radius of Bluetooth is limited to about 10 meters for the transmitters used in most of the mobile phones and that the case study in Larsen et al. (2013) is to get insights on the internal structure of a crowd, the authors analyze the data at two levels. One of them is Micro Groups Modelling.

In this level, the objective is to discover if people move alone or in groups and how groups are different In Larsen et al. (2013), micro groups are sets of people frequently co-occurring in the same area and time frame (spatio-temporal bin). After dividing the timeline of the entire festival into 1076 x 10-minute temporal bins and with each scanner creating a spatial bin, the authors started by out of the 8,534 unique devices discovered, removing the ones that were seen in less than 10 temporal bins or less than 3 spatial bins. Either because being in less than 10 temporal bins didn't provide enough data or because being in few spatial bins meant that the scanned device was a stationary one (such as crew laptops). For all common occurrences, a directed graph was constructed having the weight of each link estimated by the number of co-occurrences of the participant A with participant B, divided by total number of occurrences of the participant A (A to B edge).

This method had many constraints: from 130,000 participants, participants had to be seen in the same 10 meter radius within the same 10 minutes at least half of the times they were observed in total and in at least 3 different locations, to ensure meaningful data for the purpose. Regardless, the authors were still able to detect micro groups, having in the end 500 people moving around while belonging to a particular structure.

Macro Modelling

In this case, the objective was to combine the spatio-temporal traces with the bands schedule, in order to find out which concert each of the participants attended (Larsen et al., 2013).

After assigning meta information to each show (genre, playcount, etc), the authors first approach was to calculate the Pearson's correlation between the number of unique devices found during each concert and the logarithm of the playcount of the band. A strong correlation was expected, as the number of people at a concert seems to be correlated with how much the band is listened to (playcount). In the end, for the Larsen et al. (2013) the Pearson's correlation proved that people's choices regarding concerts are not really correlated to the band's popularity. So a more complex model was used.

In terms of pre-processing, the authors transformed the time-series data into a binary attendancy table, having a matrix that maps each participant to the concerts attended by the person. After that, they transformed it into a matrix that indicated how many times each participant was scanned at a certain concert. Finally, the table is again transformed to a binary matrix by filtering all the entries which had a lower value than a certain threshold. To remove outliers, some similar processing was done as in the Micro Modelling. Participants who participated in less than three concerts and Bluetooth devices recorded in the same location throughout the festival were removed, having remained 5127 attendees for further analysis.

Regarding the model, the data was fitted in an Infinite Relational Model to reveal the pattern's of people behaviour at the festival. The model's stability and generalizability are proven by different number of measures used to evaluate it as well as the usage of non-complete datasets and the insertion of randomness in the testing phase. Some interesting insights were obtained in this process. For example, in the image below (3) it's possible to see that the user cluster 5 is highly associated with the concert cluster 15 or 20, meaning that people in cluster 5 attended concerts from these clusters.

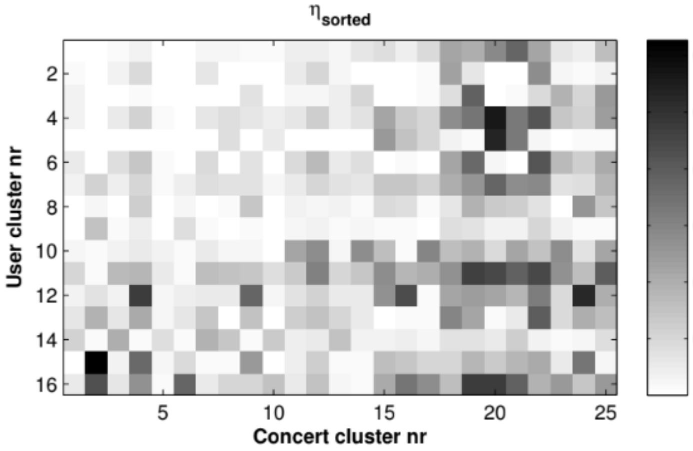


Figure 3: Figure that demonstrates the relation between user and concert clusters in Larsen et al. (2013).

Discussion and Conclusions

To sum up, this paper proved that even though Bluetooth based approaches to crowdsensing can have some drawbacks (like having to be in discoverable mode for example), it has proven in real life its feasibility. Using Bluetooth signals as a way to detect persons might not be the best solution, since these signals have a really short range of detection. Additionally, the analysis of the spatio-temporal data proved to reveal interesting insights when wanting to discoverer relations between each user that was sensed and also between the user and a concert. The issue is that nowadays some smart devices have the MAC randomization enabled, so the present approach wouldn't work so well nowadays. Nevertheless, the authors made sure to protect one's privacy by hashing the MAC address of each device. The device used was Nokia N900, which offered 3G communication, data storage, battery power and Global Positioning System tracking. Having a smartphone to essentially sense Bluetooth signals and be responsible for the storage of data is not worth the price of the device. Finally, the authors say that all the data and its insights were being displayed in a 46 inch monitor during the festival and that the participants were attracted by it, which can reflect on how people are becoming more open-minded towards these kind of technologies (Larsen et al., 2013).

2.4.2 *Sensing through Bluetooth and Wi-Fi*

In Schauer et al. (2014), the main objective is to present pratical approaches of crowdsensing, such as crowd density and pedestrian flows estimation. The methods are tested in a real life scenario. In fact, all the sensing occurs in a major German airport, which is not only beneficial because of being a real scenario, but also because the authors use ground truth information provided by the security check to test the feasibility of their approach.

Methodology

During the period of 16 days, the authors obtain data by using 2 time-synchronized laptops to discover devices with Bluetooth or Wi-Fi enabled. These two laptops were placed in two different locations: one located in the public area, and the other inside the area after the boarding pass scans - the security area. As in Larsen et al. (2013), this papers draws attention to some of the disadvantages of using Bluetooth as a way of crowdsensing. On the other hand, the wireless local area network technology, know as Wi-Fi, does present some advantages in crowdsensing when compared to Bluetooth. Firstly, instead of having a communication range of 10 meters like Bluetooth, it can go from 35 meters in indoor environments to 100 meters for outdoor environments, which facilitates the identification of Wi-Fi enabled devices. Besides that, the authors claim that in different mobile devices the active scan - which is the process that allows that detection of the Wi-Fi probes emitted by such devices - occurs at least once within two minutes.

Data

Regarding data structure, the data collected on these experiments had fields like the RSSI value, the MAC address and the entry's time stamp. In the 16 days, there were over 11 million probe requests captured in the public area and about 8.5 million in the private one. Daily, the authors were able to capture 9,995 unique Wi-Fi MAC address and 357 unique Bluetooth addresses. With this data, the Schauer et al. (2014) concluded that Apple devices send out probe requests more frequently when compared to other Android devices.

Crowd Density Estimation

The Schauer et al. (2014) define crowd density estimation as: *amount of people per unit of area within a certain time interval*. Therefore, estimating this measure in the node's coverage is done by counting the amount of unique devices during a certain period of time. In order to evaluate the results, the authors expected that whenever there was a higher frequency of board pass readings, the crowd density should be higher too. In the figure below (4), it is possible to see that the correlation exists and that there is a higher density of people in the public area, which was expected.

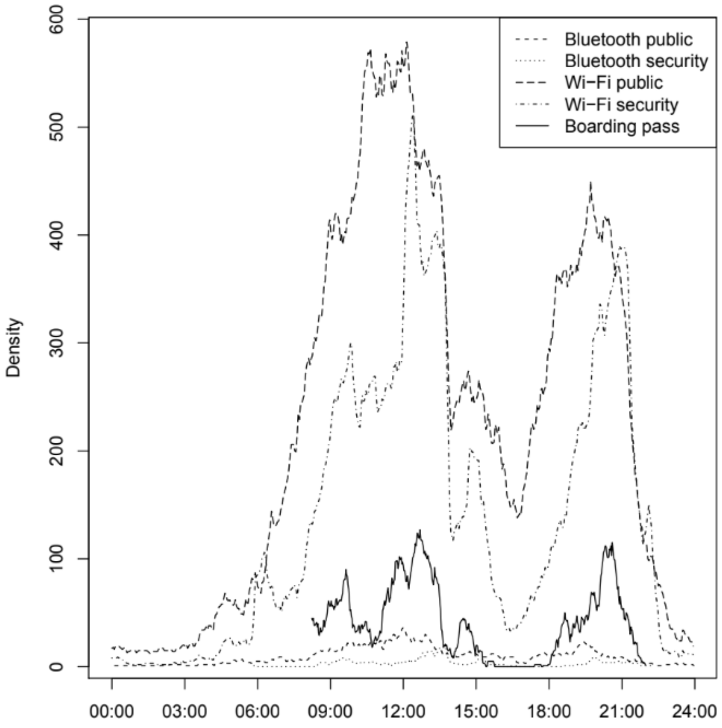


Figure 4: Daily data of Wi-Fi, Bluetooth, and boarding pass readings in Schauer et al. (2014).

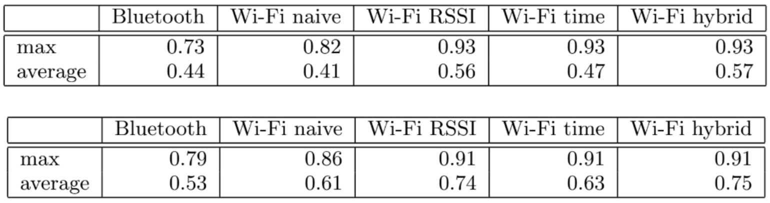
Although there was no data representing the ground truth, it is possible to conclude that the density of the crowd had some relation with the actual number of people that were on the boarding pass readings.

Pedestrain Flow Estimation

The Schauer et al. (2014) define pedestrian flow as: *amount of people moving one way through an area of interest within a certain time interval*. In order to estimate this measure, the paper presented three different approaches: Naive Approach, Time-based Approach, RSSI-based Approach and finally Hybrid Approach.

The simplest way, the Naive Approach, was to count the number of unique MAC addresses which were captured at the both sensors (s1 and s2) within a specific time interval. Two big problems arised from this method. One of them was that with this method it wasn't possible to determinate the direction of a person. The other, was that a device present in the overlapping zones of each sensor would be accounted as a person moving, increasing the rate of false-positives. The first problem was solved using a Time-based Approach, in which the specific time interval had to be positive (between the first and last node). In the second case, the authors presented a RSSI-based approach in which the pedestrian flow is calculated as the number of unique MAC addresses in a certain period of time that have the RSSI value over a certain threshold for both nodes. Similar to Larsen et al. (2013), Schauer et al. (2014) refer that the RSSI can be influenced by numerous factors, such as the environment itself or the device characteristics. Therefore, defining a reasonable threshold was the key here, but the Schauer et al. (2014) concluded that it was necessary a more of a Hybrid Approach. Finally, this method uses both the RSSI value and the time of when a MAC address was captured. This lead to the following method to calculate pedestrian flow: number of unique MAC addresses captured containing a positive time delay between the sensors and at least one capture with the RSSI value over a certain threshold for both nodes. This way, this method ensures that the sensed pedestrian is moving from one point to another, while also reducing the possibility of false-positives in case the pedestrian was in the overlapping detection zones of the nodes.

This final approach offered some good results, but in order to improve them the Schauer et al. (2014) decided to only capture data when the security gate was open. This ensured that the captured data could be compared with the ground truth having less false-positives and therefore a focused estimation. Below (5) is presented the Pearson Correlation (which indicates how much one observation is correlated to another, from a scale to 0 to 1 Schauer et al. (2014)), of each approach and also with the optimal time shift for sensoring being implemented.



	Bluetooth	Wi-Fi naive	Wi-Fi RSSI	Wi-Fi time	Wi-Fi hybrid
max	0.73	0.82	0.93	0.93	0.93
average	0.44	0.41	0.56	0.47	0.57

	Bluetooth	Wi-Fi naive	Wi-Fi RSSI	Wi-Fi time	Wi-Fi hybrid
max	0.79	0.86	0.91	0.91	0.91
average	0.53	0.61	0.74	0.63	0.75

Figure 5: Correlation coefficients for each approach and with different focused estimations in Schauer et al. (2014).

Note that Schauer et al. (2014) performed different times shifts for each estimation and then calculated the correlation coefficient for each time shift.

Discussion and Conclusions

The authors were able to prove that for crowdsensing the Wi-Fi is more reliable than Bluetooth not only because of its characteristics but also because more modern devices use it nowadays. Besides that, Schauer et al. (2014) proved that it is possible to estimate crowd density and pedestrian flow using the Wi-Fi and Bluetooth as a way sense to a certain area. Additionally, the correlation results between the sensed data and corresponding boarding pass readings, showed the feasibility of the method implemented, having an average of 0.75 in the Pearson's correlation factor with their best approach. The authors also show an interesting way to counter the overlapping coverage area of various sensors. Moreover, the RSSI value was used in the calculation of the Pedestrian Flow, but because of its volatile values the method used to counter this issue was to define a threshold as a way to filter results and combine this metric along with others. The way this threshold was defined is not explained explicitly. Regarding the sensing device used, using a cheaper option (e.g. a micro controller), even if with a higher number of sensors deployed, would have the same end result. Finally, either on the Crowd Density or Pedestrian Flow estimation, the authors claim that MAC randomization in 2014 didn't raise a problem when wanting to recognize a certain device. But, time has passed and Apple is not the only integrating MAC randomization mechanism in their devices now. Despite being proved that this mechanism is still on early stages, it should be a factor to take into account when wanting to do some sort of crowdsensing.

2.4.3 *Sensing in a SmartCity context*

In Ares et al. (2016), the main objective of the crowdsensing is to address and improve issues that could be solved in a Smart City's context. In fact, a mobility monitoring system is presented as way to improve several areas of city, like traffic or security issues inside buildings. Whilst the sensing phase applied to the 4 different areas of study is the same, the authors then use different methods for analysis, processing and modelling algorithms for the different cases.

Methodology

The sensing device used in Ares et al. (2016) is a single-board computer, based on a Rasperry Pie board. As the objective is to sense each individual through the the possesion of digital devices, the board has a Bluetooth and Wi-Fi cards, allowing the board to detect Wi-Fi and Bluetooth probe requests.

The monitoring system developed by Ares et al. (2016) has an architecture of 6 modules. The first one is the software implemented in the sensor in order to detect Bluetooth and

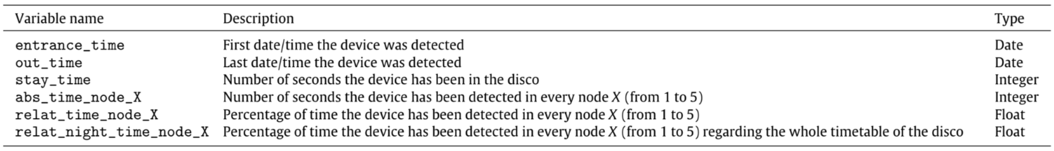
Wi-Fi devices along with extraction and encryption of the devices' MAC address. It is also responsible for publishing this data to the server. Secondly, there's a module to act as gateway between the sensors and the server, making possible the communication between the sensors themselves, and the sensors and external networks. Then, there are modules responsible for control of the sensors from distance and for the storage either in a local server or in a cloud-based storage, which have services for data mining, ML, forecasting algorithms and visualization methods.

As in any other crowdsensing case, privacy was also one of the thematics discussed by (Ares et al., 2016). In this case, the authors argue that as soon as a MAC address (either on a Wi-Fi or Bluetooth probe) is captured from a device, it is immediately encrypted using the SHA1 hash algorithm. Apart from that, the system proposed only publishes information about general data, meaning that neither individual data is shared nor data relating to a specific device.

Analysing people's mobility in a discotheque

One of the use cases that Ares et al. (2016) tested their approach in was a discotheque. A place where from common knowledge almost every one has a smartphone and where the change, in terms of mobility, is very fast. In order to collect the data, five devices were installed in a discotheque, one in each main room and the others on the main entrance and outdoor terrace. It is also mentioned - even though the authors don't expand on it - that given that fact that the eletromagnectic scenario was heavy, the testing of the device was more difficult.

In terms of amount of data, a total of 2200 different devices were detected in one of the busiest nights. After collecting the data, the main goals were to find group behavioural patterns in the data, which could optimize the decision making for marketing or security issues in the disco. To do this, Ares et al. (2016) first extracted several variables (shown in 6) from the data so that they could then apply clustering methods.



Variable name	Description	Type
`entrance_time`	First date/time the device was detected	Date
`out_time`	Last date/time the device was detected	Date
`stay_time`	Number of seconds the device has been in the disco	Integer
`abs_time_node_X`	Number of seconds the device has been detected in every node *X* (from 1 to 5)	Integer
`relat_time_node_X`	Percentage of time the device has been detected in every node *X* (from 1 to 5)	Float
`relat_night_time_node_X`	Percentage of time the device has been detected in every node *X* (from 1 to 5) regarding the whole timetable of the disco	Float

Figure 6: Variables extracted from the data in Ares et al. (2016).

The clustering method used in Ares et al. (2016) was a Self Organizing Map. This method uses a feed-forward neural network that by using an unsupervised training algorithm and nonlinear regression techniques, is able to cluster the data and finally find interesting relations between the set of variables. By transforming the resulting data of the clustering method into a Unified distance matrix, the authors generated a graph that visually represented the multi-variable dataset in a two-dimensional display. Using this U-matrix + SOM approach, several graphs were produced. For instance, one of the graphs produced by SOM was able to show that in room of the sensor 112, the Main Room, the number of devices staying there a higher percentage of time (variable relat_time_node_X) was the highest between all rooms, followed by the sensor 122, as demonstrated in the figure below (7).

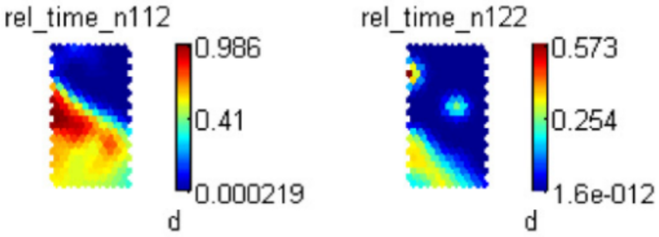


Figure 7: SOM plane analysis for the variable referring to the time a device spent in a nodes 112 and 122 in Ares et al. (2016).

Through this use case it was possible to gather and process data that can be useful for different purposes. The clustering method used showed interesting insights about people's mobility in different places inside the disco, making it possible to address issues on marketing or energetic efficiency in the context of Smart Cities (Ares et al., 2016).

Traffic Forecasting

One of the other interesting use cases in Ares et al. (2016) is regarding Traffic Forecasting. Sensing is not only useful to get insights or relevant live information about different environments, but also to forecast relevant situations like whether traffic is going to be heavy or not at certain hours in certain places.

Compared to the previous use case, in this one there was real data provided by the Spanish traffic management agency, so even though detecting cars at high speed seemed to be an issue, the validation of the approach in Ares et al. (2016) was possible by comparing the real data with the data collected by their approach. In order to collect the data, six nodes were placed in 6 different positions along different roads (that had loop detectors installed by the traffic management agency), as illustrated in the figure 8.

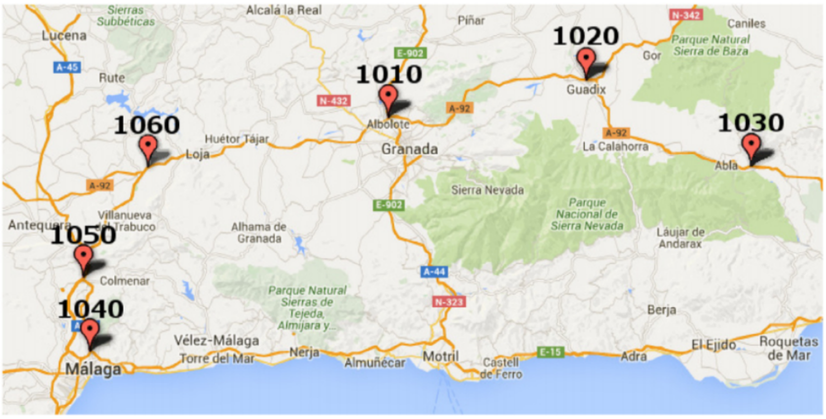


Figure 8: Location of the 6 sensors in Ares et al. (2016).

Since one of the goals in Ares et al. (2016) was to forecast the traffic, the sensed data had to be processed so that there was information about the number of cars that passed a given point in a certain period of time, organizing the data in a time series way.

Given this processing, the authors in Ares et al. (2016) tested 4 different methods (using the forecast package of the R language (Hyndman and Khandakar, 2008)) for forecasting the traffic, enabling the prediction of the number of cars that pass in different points in certain periods.

As stated in Ares et al. (2016), the method that had the most accurate behaviour was the ETS. For this experiment the data used was from the node 1010, having a total of 1920 values. The figures below (9) demonstrates that the ETS method was able to forecast the values with good accuracy when compared to the real data.

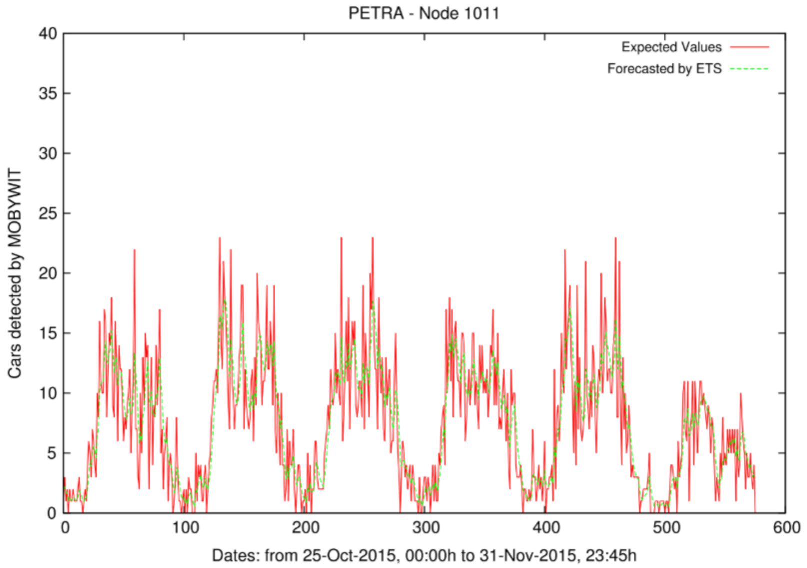


Figure 9: Expected vs forecasted number of cars in one of the highway points in Ares et al. (2016).

Finally, this use case showed that it is possible to sense data even from cars that are moving on a high speed and also suggest which forecast method might work best for time series problems. Having this embedded in a Smart City would improve allow the route planning or the improvement of traffic jams, for example.

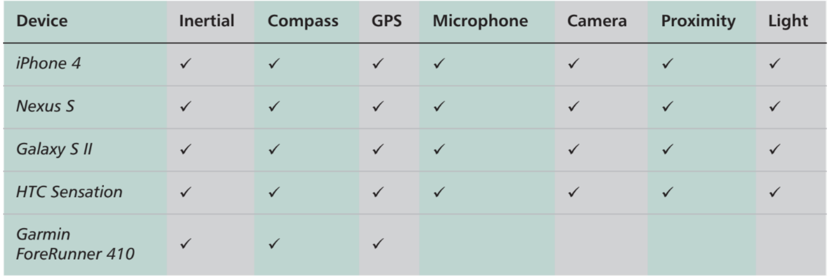
Discussion and Conclusions

The approach in Ares et al. (2016) used a singled-board computer to detect Wi-Fi and Bluetooth probe requests in order to do the crowdsensing. Privacy-wise, the authors used the SHA1 hash algorithm in order to encrypt the MAC address as soon as it was captured from a device, protecting any individual's data with this method. In the first use case, the use of clustering methods proved to be valuable when wanting to show group behavioural patterns in the extracted data. Nevertheless, in this case there were sensors distributed in various places of a discotheque, but it was not clear on whether there was overlapping of the sensors coverage area or not, and if it was something that was taken into account. A possible way to do this could be using the RSSI value to filter unwanted values. In the second use case, the proposed solution allowed the forecast of traffic related data. It was referred that vehicles travelling at high speed could be difficult to detect, but the results showed that when compared to the ground truth information provided by the loop detectors, the sensed data was accurate. Finally, MAC address randomization wasn't mentioned at all throughout the paper, which means the authors didn't consider the possible effects of this mechanism on their results.

2.4.4 *Mobile Crowdsensing*

All the approaches referred before have one point in common: they use passive sensing techniques, allowing the detection of users without them playing an explicit role in the gathering of data. Mobile Crowdsensing is a paradigm that offers a different approach, where individuals with sensing and computing devices collectively share and extract data for a common purpose (Ganti et al., 2011).

Instead of using typical IoT devices, such as RFID tags (Cook et al., 2009) or single board computers (Ganti et al., 2011), this approach defends that using mobile devices can be more useful. Commonly equipped with various sensing capabilities and also with powerful computation, as figure 10 shows, mobile devices can offer different possibilities when it comes to crowdsensing.



Device	Inertial	Compass	GPS	Microphone	Camera	Proximity	Light
iPhone 4	✓	✓	✓	✓	✓	✓	✓
Nexus S	✓	✓	✓	✓	✓	✓	✓
Galaxy S II	✓	✓	✓	✓	✓	✓	✓
HTC Sensation	✓	✓	✓	✓	✓	✓	✓
Garmin ForeRunner 410	✓	✓	✓				

Figure 10: Some of the sensing capabilities of different mobile devices in Ganti et al. (2011).

MCS has characteristics that offer not only new opportunities but also big challenges. When it comes to advantages of this paradigm, mobile devices have a bigger variety of sensing capabilities and more computing, storage and communication resources than a traditional IoT sensor. Besides that, the fact that mobiles devices are already deployed in one's life may lead to a easier usage of this devices as sensors to our everyday quotidian, improving it in several areas. Instead of using specific sensors for example to collect traffic data, a possible solution could be using smartphones carried by drivers to collect the traffic data (Ganti et al., 2011).

In terms of drawbacks, the main one is that fact that MCS, either on participatory sensing or opportunistic sensing, involves an active role of the user in order to extract any kind of data. On the one hand, participatory sensing can be defined as sensing that *requires the active involvement of individuals to contribute sensor data*. On the other hand, opportunistic is the sensing where *user involvement is minimal* (Ganti et al., 2011). In the end, either one of these approaches are inserted into the process know as Crowdsourcing, defined as *The practice of obtaining needed services or content by soliciting contributions from a crowd of people, especially from an online community* (Guo et al., 2015). More specifically, if users reported the available parking spots with text or images, that would be inserted into MCS (Ganti et al., 2011). The main issue is that the owners of mobiles devices commonly are not willing to contribute for the sensing, processing and communication of the data, unless they have an appropriate incentive mechanism to do so (Ganti et al., 2011). Besides this main disadvantage, MCS might have problems to deal with the different structure of data produced by different mobile devices (even for the same purpose), privacy related issues or the need of communication with a large number of devices (Ganti et al., 2011).

2.4.5 *Sensing in a MCS paradigm*

Following up on the MCS section, this section will present an approach embedded in this paradigm, where multiple mobile phones are used collaboratively to estimate crowd density through the scan of the environment for Bluetooth devices (Weppner and Lukowicz, 2013).

Methodology

In Weppner and Lukowicz (2013), the method proposed is to sensor the environment for Bluetooth enabled devices through a reduced numbers of users (stationary or dynamic) that are equipped with a Bluetooth scanning mobile phone to determine the crowd density in an area of $2500m^2$. A reduced number of volunteers walk through specific areas during specific times in order to extract all the relevant data. This is tested during three days at the European soccer championship official public viewing event, which has thousands of visitors. In their experimental setup, 10 students are divided in 5 teams of 2 students each, where some of them stand on the same spot (like near entrances), and others walk continuously around the event area, in order to cover all the relevant regions. In terms of the sensing device, each student is equipped with one Android smartphone, that continuously runs an applications that scans for discoverable Bluetooth devices, producing data that is saved onto a microSD card. Each scan is defined as a time interval that contains a dataset with all the unique Bluetooth devices (except the repeated devices), along with other information like the RSSI value, etc. Along with this, all the data extracted can be compared to ground truth information about crowd density with the use of a HD video camera, which can prove (or not) the feasibility of this method.

Feature Engineering

As the authors in Weppner and Lukowicz (2013) want to estimate precisely the crowd density of the whole event area, 6 features were created. In ML, Feature Engineering is the process of creating feature based on the raw data extracted, in this case from the sensors. This process has the ability to increase the performance of the prediction model, as any model needs to be fueled with relevant data, regardless of the complexity/power of the ML model (Rencberoglu, 2019; Koehrsen, 2018).

This section will explain briefly two of the features created. The first feature, *Averaged sum of distinct devices discovered by all sensors in scan window*, describes the average number of unique devices discovered in each sensor for every scan window (snap-shot), which can be seen in the figure 11. An issue that could arise from this method was that the same Bluetooth device could be detected by different sensors, which could give false results. In order to solve this, Weppner and Lukowicz (2013) argue that Bluetooth devices discovered by multiple sensors at the same time don't influence the calculation of the referred feature. One other feature created was *Ratio of discovered devices in current snapshot to discovered devices in last x minutes*. This feature allowed the creation of crowd movement insights during every snap-shot. The calculation of this feature was done by dividing the number of occurrences in the union of unique devices discovered by all sensors in a snap-shot, by the number of occurrences in unique Bluetooth devices discovered in the last 15 snap-shots. In the end, a high value of this feature would indicate that in the current snap-shot there was strong crowd movement.

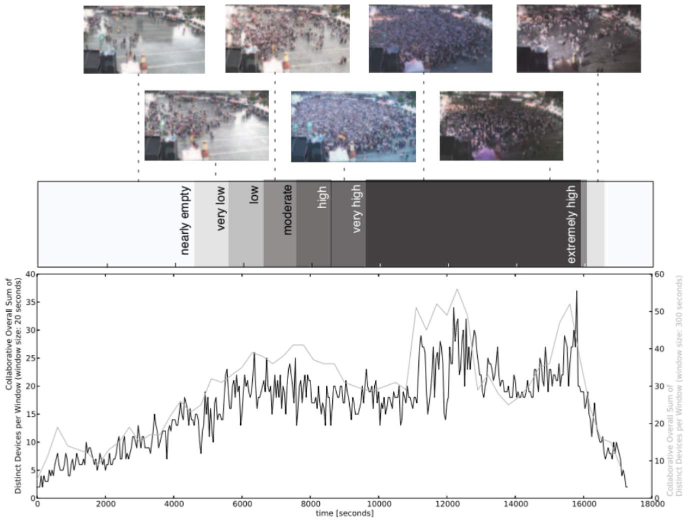
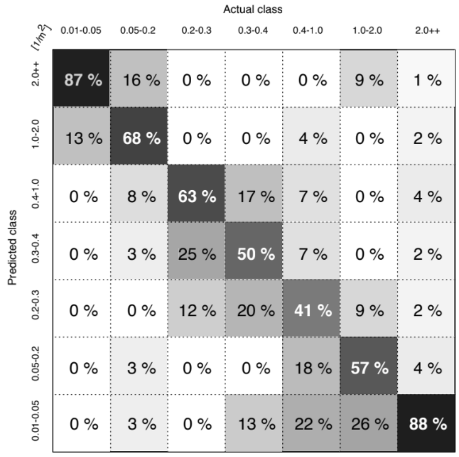


Figure 11: Feature *Averaged sum of distinct devices discovered by all sensors in scan window* along with ground truth crowd density levels in Weppner and Lukowicz (2013).

Model Evaluation

After the sensing and the processing of the data, Weppner and Lukowicz (2013) used a decision tree classifier for classifying the crowd density in 6 different levels: empty (0.01 - 0.05people/m^2), very low (0.05 - 0.2people/m^2), low (0.2 - 0.3people/m^2), moderate (0.3 - 0.4people/m^2), high (0.4 - 1.0people/m^2), very high (1.0 - 0.05people/m^2) and extremely high (2.0 + people/m^2). The evaluation was done using a 10-fold cross validation. The results for each level are illustrated in the figure below (12).



Actual class

Predicted class (1/m²)	0.01-0.05	0.05-0.2	0.2-0.3	0.3-0.4	0.4-1.0	1.0-2.0	2.0++
2.0++	87 %	16 %	0 %	0 %	0 %	9 %	1 %
1.0-2.0	13 %	68 %	0 %	0 %	4 %	0 %	2 %
0.4-1.0	0 %	8 %	63 %	17 %	7 %	0 %	4 %
0.3-0.4	0 %	3 %	25 %	50 %	7 %	0 %	2 %
0.2-0.3	0 %	0 %	12 %	20 %	41 %	9 %	2 %
0.05-0.2	0 %	3 %	0 %	0 %	18 %	57 %	4 %
0.01-0.05	0 %	3 %	0 %	13 %	22 %	26 %	88 %

Figure 12: Confusion matrix generated by the decision tree classifier in Weppner and Lukowicz (2013).

An accuracy of 75.3% was achieved, which, if the ground truth data was actually accurate, does provide relevant predictions when it comes to crowd density. But Weppner and Lukowicz (2013) does say that the ground truth information might be noisy, which doesn't contribute for an actual feasibility of the approach and model used. Besides this, Weppner and Lukowicz (2013) also argue that the human body has a high absorption coefficient of the Bluetooth signal, which affects the scanning done by the mobile devices and therefore leading to results that might not be the pretended. Relying on features that are not directly dependent on the absolute number of discoverable devices and using relative features based on the ratio between values observed by different devices, lead to improvements, which is a point to be considered. This process, feature engineering, not only helped on improving the ML model itself, but also with the visualization of meaningful data and to a more robust system overall. Nonetheless, the authors refer that there was a 30% improvement in accuracy when using multiple sensors, compared to using a single device, but don't demonstrate it. Since the presented approach is part of a bigger paradigm, know as participatory sensing, it would have been relevant to know how the volunteers who were using sensors were convinced to be part of the sensing. Lastly, their current approach would have to suffer some changes in order to have the same success, since the MAC address randomization is more present nowadays than it was in 2013.

2.5 MACHINE LEARNING

Machine Learning (ML) can be defined as a field of Artificial Intelligence (AI) (see picture 13) where computers are programmed to learn from data (Géron, 2017; Shalev-Shwartz and Ben-David, 2014). The main advantage of ML over a human's learning is that a computer has the ability to consume huge amounts of data and detect and analyze its patterns that are outside of the human perception (Shalev-Shwartz and Ben-David, 2014). This technological method's evolution arised from the increased data collection and ML algorithm's complexity, as well as the decreased cost of data processing power (Mitchell, 1991). Nowadays, ML is used in many different cases like voice/face recognition, recommendation systems or classification problems. For each kind of problem, there's a different kind of ML algorithms that normally perform better and belong to one of the four defined ML Paradigms: **Supervised, Unsupervised, Semisupervised** and finally **Reinforcement learning** (Géron, 2017).

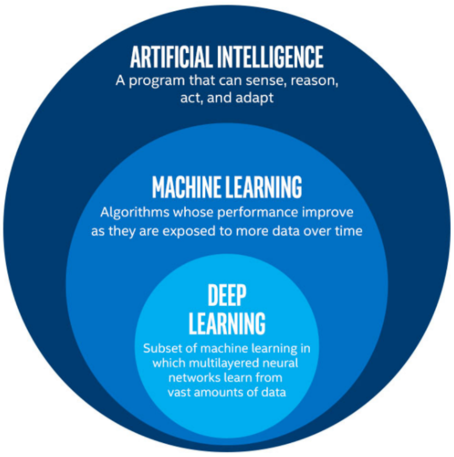

ARTIFICIAL INTELLIGENCE
A program that can sense, reason, act, and adapt
MACHINE LEARNING
Algorithms whose performance improve as they are exposed to more data over time
DEEP LEARNING
Subset of machine learning in which multilayered neural networks learn from vast amounts of data

Figure 13: Figure to demonstrate the difference between ML and AI in Singh.

2.5.1 *Learning Paradigms*

ML Paradigms vary according to each type of problem which in their turn have different amounts and types of supervision in training.

Supervised Learning

In the case of Supervised Learning, the data used to train the model contains labels, which are the outputs for each input. The model has a mapping function, which is formed by a algorithm (which differs depending on the kind of problem), that after trained predicts the output data for each input. (Géron, 2017; Brownlee)

A typical example is a classification problem, where the model needs to predict the correct categorical output, by calculating whether an email is spam or not. Given that the training the dataset contains each email labeled, for each email (input), the model must learn from the training set how to predict the label (Spam or Not Spam), in the most accurate way possible for each new email. Another kind of Supervised Learning problem, is the regression problem which happens when a certain number (a discrete output), like the price of a house for example, has to predicted by the model given inputs, in this case, containing information like number of rooms of each house, etc (Géron, 2017; Shalev-Shwartz and Ben-David, 2014).

Unsupervised Learning

Secondly, in Unsupervised Learning the difference is that the data used to train the model is unlabeled. The model now tries to learn without any supervision. In this case, the model only has available the inputs without any corresponding outputs, having to discover interesting patterns in the data (Géron, 2017; Brownlee). This kind of problem can be divided into main three categories: Clustering, Association rule learning and Visualization and dimensionality reduction. In the first place, Clustering is grouping up each input such that each one ends in the same cluster; in the same group. Next, Association rule learning is having to pair up groups of data, so that it forms a rule such as people who see the movie X also see the movie Y. Finally, Visualization and dimensionality reduction, in the case of Visualization is presenting the data so that it is understandable how the data is organized. In the case of dimensionality reduction, it is discovering which inputs reflect better the general dataset, simplifying the data without losing information. (Géron, 2017; Shalev-Shwartz and Ben-David, 2014; Brownlee)

Reinforcement Learning

Finally, Reinforcement Learning uses a method based on reward/penalty of each action to train the model. The learning system, called an agent in this context, must learn how to choose the actions that can get the most reward over time, based on its policy. In fact, the policy of the learning system is what defines its behaviour to different situations (Géron, 2017).

One of the most well know examples of the use of this paradigm is the AlphaGo. By implementing Reinforcement Learning algorithms, the robot or program learned how to play the game Go by analyzing millions of games and playing games against itself, which allowed the robot to get better over time and eventually beat the world champion Lee Sedol (Géron, 2017).

2.5.2 *Learning Algorithms*

Embedded in each learning paradigm there are a variety of algorithms used for solving typical ML tasks. Below, some of these algorithms are going to be described.

Linear and Logistic Regression

In the first place, Linear Regression is a ML algorithm inserted into the **Supervised Learning** paradigm. Being used to predict numerical values, this algorithm prediction method is to compute a weighted sum of the input variables (typically features given in a dataset) plus a bias term, which can help the model to make more accurate predictions. In the second place, Logistic Regression also belongs to the same paradigm referred before, but it is used for classification problems, meaning that it is used to predict categorical outputs. Using a similar equation as Linear Regression, this algorithm outputs the logistic of this result, showing how likely an instance is to belong to a certain class (output variable) (Géron, 2017).

Decision Tree and Random Forest

Decision Tree is a **Supervised Learning** versatile algorithm that is able to perform both classification and regression tasks. This algorithm's response is based on the criteria defined by each node of the tree. Given a certain input, each node will be responsible for decisions that will lead to the final prediction, present in the tree's leaves. Despite being a powerful and versatile technique, it might encounter over-fitting problems (Géron, 2017; Ray, 2019). In fact, the Random Forest algorithm, also embedded in the same ML paradigm, is one of the solutions used to overcome this issue. Random Forest can be defined as an ensemble of Decision Trees. Various Decision Trees are trained each on a random subset of the training set. This way, the model makes predictions based on various Decision Trees, being more likely to be more flexible with less bias and variance (Géron, 2017; Desai, 2020).

Principal Component Analysis

The Principal Component Analysis, most know as PCA, is an **Unsupervised Learning** algorithm used for extracting the most relevant features in a dataset. A dataset might have a large number of features, what PCA does is to find the ones who preserve the maximum amount of variance. In fact, it allows the to turn an intractable problem into a tractable one, without losing much information (Géron, 2017; Desai, 2020).

K-means

The K-means is an **Unsupervised Learning** algorithm, which purpose is to form clusters - groups of data that were formed because of the similarity between its data points - to discover underlying patterns in the data. After manually defining the number of clusters that are going to be created, this algorithm proceeds to optimize the position of the centroids (location that represents the center of the cluster), so that it successfully creates the intended groups of data (Desai, 2020; Garbade, 2018).

Artifical Neural Networks

The Artificial Neural Networks are the core of Deep Learning. Compared to other algorithms, the ANNs offer new possibilities to deal with highly demanding ML tasks as they are very powerful, versatile and scalable (Géron, 2017). Shortly, ANNs can be defined as a directed graph with neurons instead of nodes and links instead of edges (Shalev-Shwartz and Ben-David, 2014). The neurons are responsible for performing some sort of calculation and the result of this calculation will be multiplied by a weight as it travels through the network. These components are present in the layers of a network including the hidden layers, where the calculation happens, the input layer - which contains the data provided to the ANNs - and also the output layer - responsible for producing the outputs (Josh, 2015). There are various types of ANNs, including the Convolutional Neural Network, Recurrent Neural Network or Long Short-Term Memory, for example. Each one has its own advantages and utility in different use cases.

2.6 CONCLUSIONS

After the analysis of several articles, it was possible to identify that crowdsensing can be done in numerous ways and for different purposes, as illustrated in table 1. In general, from the reviewed the articles, even though some of them belonged to different fields of study, the main approach was to capture either Wi-Fi or Bluetooth probe requests from smart devices as a way to sense the crowd. With the extracted data, the articles showed that is possible to implement crowdsensing for different purposes: crowd density estimation, traffic forecasting or analysing people's mobility are just a few examples.

	Device	**Field of Study**	**Techonologies**	**RSSI**	**#rawdataset(∼)**	**Models**	**Privacy**
Sensing through Bluetooth	Nokia900	Music Festival	Bluetooth	No	1 million	Graphs Infinite Relational Model	MAC encryption
Sensing through Bluetooth and Wi-Fi	Laptop	Airport	Wifi Bluetooth	Yes	11 million	No information available	No
Sensing in a SmartCity context	SingleBoardComputer	People's mobility in a discotheque Traffic forecasting	Wifi Bluetooth	No	No information available	SOM ETS ARIMA Theta Infinite Relational model	MAC encryption
Sensing in a MCS paradigm	SmartPhones	Public viewing event	Bluetooth	Yes	4100	Tree Classifier	No

Table 1: Crowdsensing Approaches

When it comes to compare Bluetooth based approaches to Wi-Fi one's, it was clear that capturing Bluetooth probe requests presented a higher number of challenges. Firstly, despite the fact that either Bluetooth or Wi-Fi need to be manually turned on by the user for enabling the retrieval of data by the sensor, Bluetooth has to be additionally set to discoverable mode as well. Moreover, the detection's range of Wi-Fi signals is much higher than Bluetooth ones. While the first has a communication range that can vary from 35 meters to 100 meters (depending on the environment), the second has a range of about 10 meters. Finally, in one of the articles reviewed, where their approach was tested on a big airport, the authors in Schauer et al. (2014) revealed that they detected 6,211 unique vs Wi-Fi MAC 250 unique Bluetooth addresses per day in the public area, which verifies that fact that Bluetooth probe requests are harder to capture, and therefore using Wi-Fi is a more efficient approach in this cases.

As expected, in order to do any kind of sensing it is necessary to have a sensing device. In the reviewed approaches, different kind of devices were used: Nokia N900 smartphones, laptops, single-board computers or android smartphones in general. It was possible to identify that even though these devices have different capabilities, the main one is being able to capture probe requests. Therefore, the conclusion for this dissertation is that the sensing device used should be capable of doing so. Other aspects are import too, namely the price of the device or for the capacity to connect itself to other networks.

In all the captures done by these devices the structure of data extracted has a typical set of values: Bluetooth / Wi-Fi MAC address, scanner id, timestamp, scanner id (in case multiple scanners are used) and RSSI. From this fields, the RSSI is the one that is more controversial. Since it is a very volatile value, some approaches don't use it at all. The one's who do make sure not to take any conclusions that rely solely on this metric. Defining a way to filter its values (with a defined threshold), combining it with other metrics, or creating features that despite being relevant, are not directly dependent on the RSSI, are some of the methods use to mitigate the volatileness of this metric.

Regarding the MAC address, some of the crowdsensing approaches decided to hash this address right after extracting it, in order to secure one's privacy. Crowdsensing provides solutions for modern day problems, but at the same time it handles data that, if not manipulated correctly, might bring privacy issues. Therefore, either encrypting the MAC address or being careful to only show data that represents a big group of people are some of the measures that help protecting one's privacy. A big subject arises from this topic: MAC randomization - a solution that has been recently implemented in this field. Some studies have revealed that either there are ways to get around it, or that many devices do not implement it at all. When wanting to do some sort of crowdsensing, this mechanism should not be ignored, since in spite of protecting one's privacy, it might produce data that leads to wrong conclusions. If a certain device is detected by sensors with the MAC address X, and 1 minute later it changes its address, then, the sensors would produce data indicating that apparently two different devices were detected within 1 minute, which is not true. This is not a common case to happen, since MAC randomization does have flaws, and when occurring efficiently it occurs when the smart device is sending probes in order to get search for available networks.

Another important point from the examined methods is that some of them have ground truth information, giving them the possibility of testing the output of their approach with data that corresponds to the truth. Finding a point of interest where ground truth information is available to test the approach that will be presented in this dissertation, is a key factor to test its feasibility. Either resorting to other sensing devices, like video cameras, or to places where to get in there's needed some sort of check-in method like an airport, were some of the methods used by the approaches reviewed.

In addition, the user's role in the sensing phase must always be recognized. Having an active role, like in opportunistic or participatory sensing, brings challenges that are difficult to overcome. The Mobile CrowdSensing section showed that there is the need to have an appropriate incentive mechanism so that the extraction of data is possible.

After the sensing phase, each studied work showed how they operated with the gathered data in order to make it useful and beneficial. In terms of data modeling, the studies revealed that creating new variables with the existent data, a process know as *Feature Engineering*, can be beneficial not only for visualization purposes, but also for the fueling of prediction algorithms, making them have increased robustness and accuracy. The use of different algorithms proved to be useful for various purposes: traffic forecasting, pedestrian flow estimation or even for finding relations between each user and the groups they were moving in.

3

THE PROBLEM AND ITS CHALLENGES

The purpose of this section is to explain in which areas the proposed approach can have a positive impact. Firstly, the problems that can be solved are going to be discussed. Secondly, the main expected challenges are going to be presented.

3.1 PROBLEM

The problem can be divided in two main points. On one hand, there are areas of our life that can be improved through today's technology, namely the one's related to AmI and Smart Cities. On the other hand, the studied crowdsensing solutions have revealed that there were some issues that could be fixed or improved.

AmI solutions have proved to be a great asset to improve several areas of our daily life. For example, in Urban Security these kind of solutions can improve the safety of the VRUs. It is a known fact that the number of traffic deaths is increasing every year or that there could be more information available when wanting to plan a trip according to the traffic peaks in certain places. Another example is in the area of Marketing. The decision making could be optimized if there was information about the group behavioural patterns in a certain place, making it possible to enhance each marketing decision. In either of these cases, an AmI solution, in this case crowdsensing, can be a solution to help improving these areas and several others. Ares et al. (2016) proved that crowdsensing provided insights about people's mobility in different places inside a disco, which could help enhancing security, marketing or energetic efficiency issues.

Another category of the problem is related to the crowdsensing itself. The studied articles revealed that many of the approaches not always used the most efficient solution, either in terms of cost or of accuracy, and also that there could be room for improvement in other aspects as well. The proposed solution will try to address this issues and develop a crowdsensing solution that takes into account factors like privacy, the type of sensing device, technologies used and the user role in the sensing phase, just to mention a few.

3.2 CHALLENGES

The challenges can differ from AmI or Mobile and Pervasive Computing to more specific and technical one's. One major challenge is that even though a AmI system has obvious benefits by continuously improving one's quality of life, the possibility of performing wrong actions demands caution when deploying such a system. Furthermore, the purpose of this dissertation is to develop a system that can bring advantages to one's life without any active participation from the user. Therefore, another challenge is being able to build a technology capable of shortening the gap between humans and computing devices. Moreover, from a technical perspective, both sensing and reasoning phase have challenges that must be overcome in order to develop a viable system. Starting by the sensing phase, in the case of RSSI, a metric that can be useful for calculating how far a person is from the sensor, it might be difficult to rely solely on it because of its high sensibility to obstacles in the way of the signal. Additionally, the recently implemented MAC Address Randomization forces extra testing to conclude how much a certain MAC address is a viable identifier of a device or not. Besides that, choosing a point of interest where ground truth is available might be difficult to find, but it is necessary to prove if the system is viable or not. Also, after the sensing phase, it will be crucial to get the best out of the data available: creating new relevant feature - a process called *Feature Engineering*, removing unnecessary data or choosing and testing suitable ML models are some steps to be taken.

4

DEVELOPMENT

This section will explain, in detail, the main phases of this project, all the work done so far and, additionally, identify why and which device was chosen for the purpose of this dissertation and which software was used to program it.

4.1 PROPOSED APPROACH

This project can be divided in two main phases. The first is data acquisition. This sensing will be done with the use of a Crowd Sensing Smart Scanner, more specifically a second generation ESP8266 ESP-12E NodeMCU Amica board. To program this board, The Arduino Integrated Development Environment (IDE) will be used to develop software that allows the passive detection of the Wi-Fi signals from people's smart devices. In terms of data storage, the sensor will be storing data in the Firebase Realtime Database, a cloud-hosted NoSQL database that allows the storage and synchronization of the Crowd Sensor data in realtime. With this data, the data processing will enable the estimation of a person's distance to the Crowd Sensor through the RSSI value of each occurrence.

The second phase main objective is to know the density of people in a certain area as well as the use of algorithms for crowd detection and control. This project will follow the Cross-Industry Standard Process for Data Mining (CRISP-DM) which provides an overview of the life cycle of a data mining/machine learning project. The phases are Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment (shown in picture 14). More importantly, the key aspect of this life cycle is that the sequence of the phases is not rigid, so this phases are parts of an ongoing cycle of analytics activity, having to go back forth among these phases frequently (Chapman et al., 2000).

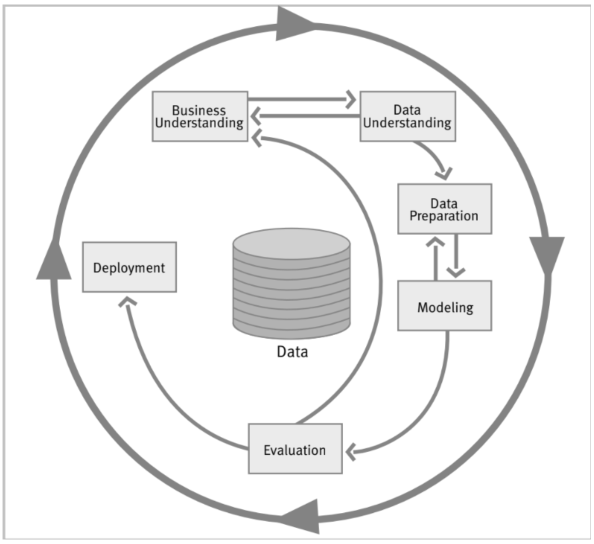


Figure 14: CRISP-DM Phases in Chapman et al. (2000)

Finally, regarding the tools that will be used after the data acquisition, the programming language Python has a lot of useful libraries. For the Data Understanding and Preparation, libraries like Pandas, Numpy or Matplotlib offer a range of different tools to process the data. As for the Modeling, it is planned to use Machine Learning and Deep Learning libraries like Keras, TensorFlow and Scikit-Learn.

4.2 DEVICE

The first step in the solution was to find a proper device for the extraction of data. Since this work is being done in the context of Smart Cities, the device was chosen taking into account the price, the capacity of detecting Wi-Fi probe requests and the capacity of uploading the sensed data to real time data base.

Taking into account these parameters, the device choosed was the low-cost Wi-Fi microchip ESP8266 ESP-12E NodeMCU Amica board (figure 15. The NodeMCU is a low-power Arduino type board which runs on ESP8266 Wi-Fi module. In addition to the reduced cost of €2.4 per board, it also has a Wi-Fi module, 4MB flash memory, a built-in antenna, open-source support, a micro-USB interface and finally small dimensions (4.8x2.4x0.5cm) and low weight (109g). Such features allow the board to be a sensor prepared and suitable for crowdsensing (Fernandes et al., 2018).

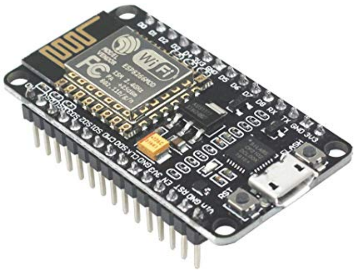


Figure 15: ESP8266 ESP-12E NodeMCU Amica board

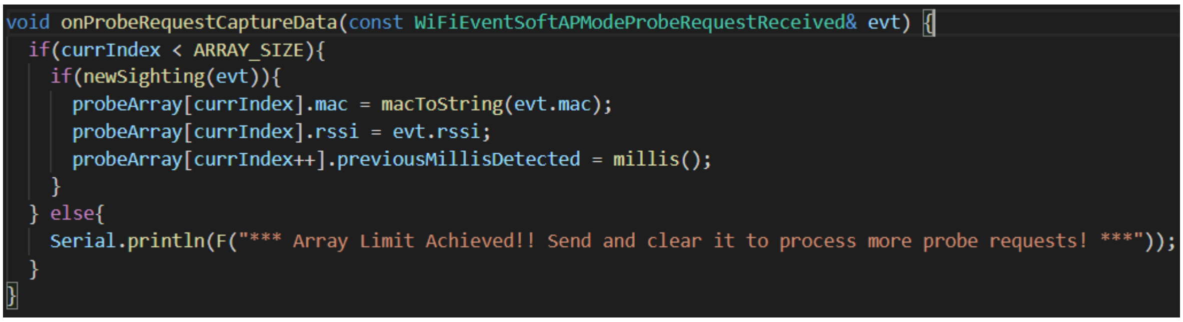
4.3 ARDUINO IDE

In order to program the board, the application that will be used is The Arduino Integrated Development Environment (IDE). It provides simple one-click mechanisms to compile and upload a program to Arduino board. This program, called a sketch in this IDE, can be written in the language C or C++. It needs to have implemented at least two basic functions: setup() and loop(). The first one is a function that is called when the sketch starts after power-up or reset of the board, being normally used for initializing variables, setting up connections or libraries needed in the developed sketch. The second is a function repeatedly executed until the board is powered of or reseted.

The Arduino project's open-source nature has allowed the development of many free software libraries that developers are using to extend their projects. This allowed the use of two external libraries in the development of the code: ESP8266Wifi (Github, a; Grokhotkov; Github, b) and Firebase Arduino (Github, c). Its usage enables the exploration of the board's Wi-Fi module, allowing the connection to a network and also the detection of Wi-Fi probe requests and additionally the possibility of uploading the data into the Firebase Realtime Database (Google, a,b).

4.4 WORK SO FAR

A sketch, developed in the Arduino IDE, was coded in a way that the ESP8266 ESP-12E NodeMCU Amica board captured Wi-Fi probe requests emitted by people's smart-devices in a non-invasive manner. The libraries referred in the previous section, ESP8266WiFi.h and FirebaseArduino.h, provide all the necessary routines to fulfil the purpose of the sketch. The *setup()* function, executed once the board is initialized or reseted, is used for setting up the board as an access point, connecting it to a network, and if connected successfully, establishing a connection to the Firebase. Furthermore, a timer is also initialized for managing the transmission of data to the Firebase database. Finally, four different handlers are registered. These handlers are called once a station connects or disconnects to the board and when a Wi-Fi probe request is captured by the board. In fact, the most important function, *onProbeRequestCaptureData()*, is the one being called upon the reception of a Wi-Fi probe request (figure 16). If the data structure used for storing the data temporarily has space, the new probe request will be stored in case of it being considered a new sighting. After that, every *sendtimer* seconds all the data is pushed into the Firebase database and cleaned locally. Apart from this, the *loop* function is also enabling the user to send commands through the serial like *Stop/Restart*, *Count*, *Start/Stop_Timer* and *Clear*, among others.

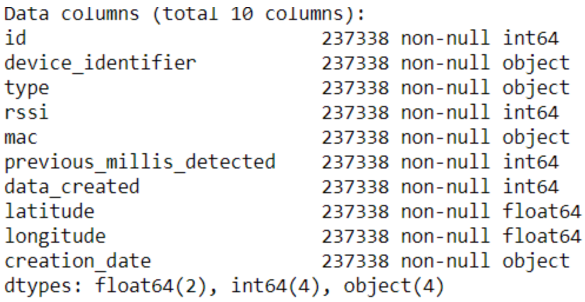


```
void onProbeRequestCaptureData(const WiFiEventSoftAPModeProbeRequestReceived& evt) {
  if(currIndex < ARRAY_SIZE){
    if(newSighting(evt)){
      probeArray[currIndex].mac = macToString(evt.mac);
      probeArray[currIndex].rssi = evt.rssi;
      probeArray[currIndex++].previousMillisDetected = millis();
    }
  } else{
    Serial.println(F("*** Array Limit Achieved!! Send and clear it to process more probe requests! ***"));
  }
}
```

Figure 16: Function called upon the reception of a Wi-Fi probe request

Besides developing the sketch explained above, a quick analysis of data was also made using a dataset produced by a similar sketch. The data's structure should be similar or equivalent to the one that's going to be produced by the proposed solution. All the analysis was done using the Pandas library from the programming language Python and the language itself.

The dataset in question has 237338 entries, characterized by 10 different attributes: id, device_identifier, type, rsi, mac, previous_millis_detected, data_created, latitude, longiude and creation_date. As demonstrated in the picture 17, in these attributes there three different types of data: *float*, *int* and *object*. Besides that, there are no missing values meaning that there's no need to any kind of data manipulation for this specific case.



```
Data columns (total 10 columns):
id                          237338 non-null int64
device_identifier           237338 non-null object
type                        237338 non-null object
rssi                        237338 non-null int64
mac                         237338 non-null object
previous_millis_detected    237338 non-null int64
data_created                237338 non-null int64
latitude                    237338 non-null float64
longitude                   237338 non-null float64
creation_date               237338 non-null object
dtypes: float64(2), int64(4), object(4)
```

Figure 17: Information about the 10 attributes of the dataset

5

CONCLUSION

5.1 CONCLUSIONS

The majority of the work done so far was focused on investigating the current paradigm of crowdsensing. The State of Art section was important mostly to perceive that crowdsensing can be done in different number of ways and that is has challenges that me be overcome to have a successful end product. It provided several insights that will be taken into account in this project. Besides that, the Wi-Fi Probing section was necessary to understand many users currently use this technology and how it evolved during the last years, specially its evolution in terms of privacy. Additionally, knowing how Smart Cities and AmI are integrated with each other and how they can be useful to our daily life through solutions like crowdsensing, was also important to prove that nowadays technology is a growing area with a lot of opportunities to make a positive impact in our daily lives. Finally, the development section it's naturally an on going work, but the work done so far proved that is possible to capture Wi-Fi probe requests using the proposed solution. With this, the solution allows the identification of a person if he or she possesses some kind of smart device that possesses Wi-Fi.

5.2 FUTURE WORK

Having developed the sensing software, the next step is to collect more data. To do so, several points of interest will be chosen so that the collection and subsequent data processing and analysis are possible. The objectives are to create relevant features to the case in question, calculate the distance between a person and the sensor and analysing the data so that the extrapolation to get insights like group behavioural data is possible. Being able to predict whether a certain place is going to be crowded or not is also a goal of this dissertation, so testing several ML models and choosing the one with better performance also has high relevance. Finally, all the work done should be explained in the writing of the thesis.

The figure below (18) demonstrates the work plan that is being undertaken in this project.

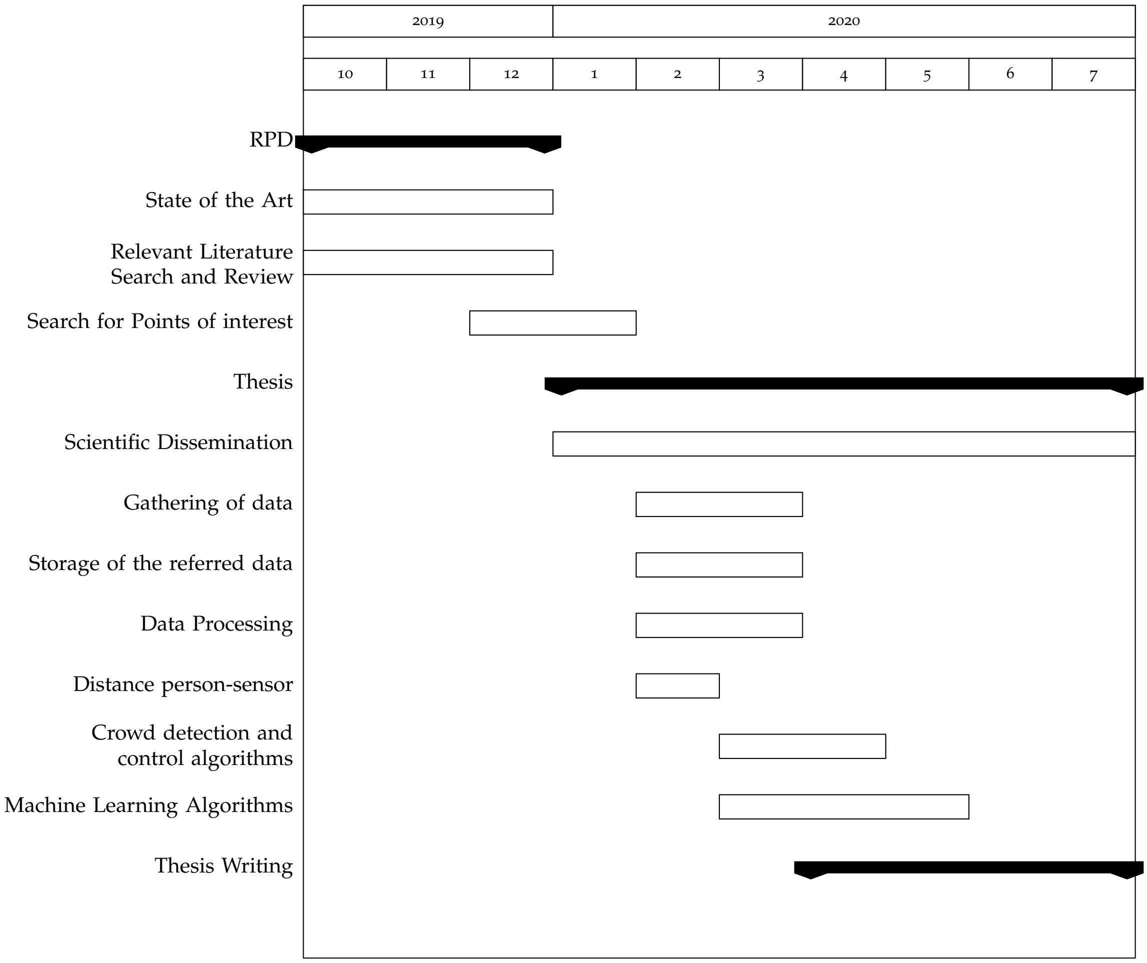


Figure 18: Gantt Diagram Chart

BIBLIOGRAPHY

A. F. Ares, A. Mora, M. G. Arenas, P. García-Sánchez, , G. Romero, V. R. Santos, P. A. Castillo, and J. M. Guervós. Studying real traffic and mobility scenarios for a smart city using a new monitoring and tracking system. *Future Generation Computer Systems*, 76, 11 2016.

M. Barbera, A. Epasto, A. Mei, V. Perta, and J. Stefa. Signals from the crowd: Uncovering social relationships through smartphone probes. pages 265–276, 10 2013.

A. Beresford and F. Stajano. Mix zones: User privacy in location-aware services. pages 127 – 131, 04 2004.

J. Brownlee. Supervised and unsupervised machine learning algorithms. URL https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/.

J. C.Augusto, H. Nakashima, and H. Aghajan. *Handbook of Ambient Intelligence and Smart Environments*, chapter Ambient Intelligence and Smart Environments: A State of the Art. Springer, Boston, MA, 2010.

P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth. *CRISP-DM 1.0*. 2000.

D. J. Cook, J. C. Augusto, and V. R. Jakkula. Ambient intelligence: Technologies, applications, and opportunities. *Pervasive and Mobile Computing*, 5(4):277—-298, 2009.

R. Desai. Top 9 algorithms for a machine learning beginner, 2020. URL https://towardsdatascience.com/top-10-algorithms-for-machine-learning-beginners-149374935f3c.

E. P. C. European Union. Directive 2010/40/eu of the european parliament and of the council of 7 july 2010 on the framework for the deployment of intelligent transport systems in the field of road transport and for interfaces with other modes of transport.

E.Longo, A. Redondi, and M. Cesana. Pairing wi-fi and bluetooth mac addresses through passive packet capture. pages 1–4, 06 2018.

M. Ermes, J. Parkka, J. Mantyjarvi, and I. Korhonen. Detection of daily activities and sports with wearable sensors in controlled and uncontrolled contitions. *IEEE Transactions on Information Technology in Biomedicine*, 12:20–26, 2008.

B. Fernandes, F. Silva, C. Analide, and J. Neves. Crowd sensing for urban security in smart cities. *Journal of Universal Computer Science*, 24(3):302–321, 2018.

J. Freudiger. How talkative is your mobile device? pages 1–6, 06 2015.

R. Ganti, F. Ye, and H. Lei. Mobile crowd sensing: Current state and future challenges. *IEEE Communications Magazine*, 49:32–39, 11 2011.

M. J. Garbade. Understanding k-means clustering in machine learning, 2018. URL https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1.

G.Delapierre, H.Grange, B.Chambaz, and L.Destannes. Polymer-based capacitive humidity sensor: characteristics and experimental results. *Sensors and Actuators*, 4:97–104, 1983.

Github, a. URL https://github.com/esp8266/Arduino/tree/master/libraries/ESP8266WiFi.

Github, b. URL https://github.com/esp8266/Arduino/blob/2.6.3/doc/esp8266wifi/readme.rst.

Github, c. URL https://github.com/FirebaseExtended/firebase-arduino/tree/master/examples/FirebaseDemo_ESP8266.

Google, a. URL https://firebase.google.com/.

Google, b. URL https://firebase.google.com/docs/database.

B. Greenstein, D. Mccoy, J. Pang, T. Kohno, S. Seshan, and D. Wetherall. Improving wireless privacy with an identifier-free link layer protocol. pages 40–53, 01 2008.

I. Grokhotkov. URL https://arduino-esp8266.readthedocs.io/en/2.6.3/esp8266wifi/readme.html.

B. Guo, Z. Wang, Z. Yu, Y. Wang, N Y. Yen, R. Huang, , and X. Zhou. Mobile crowd sensing and computing: The review of an emerging human-powered sensing paradigm. *ACM Computing Surveys*, 48, 08 2015.

Aurélien Géron. *Hands on Machine Learning with Scikit Learn and TensorFlow*. O'Reilly Media, 2017.

C. Harrison, B. Eckman, R. Hamilton, P. Hartswick, J. Kalagnanam, J.Paraszczak, and P. Williams. Foundations for smarter cities. *IBM J. RES. DEV.*, 54(4), 2010.

A. Holst. Number of smartphone users worldwide from 2016 to 2021, 2019. URL https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/.

R. Hyndman and Y. Khandakar. Automatic time series forecasting: The forecast package for r. *Journal of Statistical Software*, 26, 07 2008.

Josh. Everything you need to know about artificial neural networks, 2015. URL https://medium.com/technology-invention-and-more/everything-you-need-to-know-about-artificial-neural-networks-57fac18245a1.

J. J. Jung and A. Muñoz. Intelligent services for smart cities. *Journal of Universal Computer Science*, 24(3):246–248, 2018.

W. Koehrsen. Feature engineering: What powers machine learning, 2018. URL https://towardsdatascience.com/feature-engineering-what-powers-machine-learning-93ab191bcc2d.

J. E. Larsen, P. Sapiezynski, A. Stopczynski, M. Mørup, and R. Theodorsen. Crowds, bluetooth and rock'n'roll: Understanding music festival participant behavior. 2013.

L. Li. Technology designed to combat fakes in the global supply chain. *Business Horizons*, 56: 167–177, 2013.

L.Oliveira, D. Schneider, J. Souza, and W.Shen. Mobile device detection through wifi probe request analysis. *IEEE Access*, PP:1–1, 06 2019.

J. Martin, T. Mayberryand C. Donahue, L. Foppe, L. Brown, C. Riggins, E. Rye, and D. Brown. A study of mac address randomization in mobile devices and when it fails. *Proceedings on Privacy Enhancing Technologies*, 2017, 03 2017.

Tom M. Mitchell. Machine learning and data mining. *Communications of the ACM*, 42(11), 1991.

M.Murphy. Cellphones now outnumber the world's population, 2019. URL https://qz.com/1608103/there-are-now-more-cellphones-than-people-in-the-world/.

M.C. Mozer. *Smart Environments: Technology, Protocols, and Applications*, chapter Lessons from an Adaptive Home, pages 273–298. Wiley, 2004.

P. L. O'Connell. Skorea's high-tech utopia, where everything is observed, 2005. URL https://www.nytimes.com/2005/10/05/technology/techspecial/koreas-hightech-utopia-where-everything-is-observed.html.

Á. Petkovics, V. Simon, I. Gódor, and B. Bence. Crowdsensing solutions in smart cities towards a networked society. *EAI Endorsed Transacitons on Internet of Things*, 1, 10 2015.

E. Qin, Y. Long, C. Zhang, and L. Huang. Cloud computing and the internet of things: Technology innovation in automobile service. *LNCS*, 8017:173 – 180, 2013.

C. Ramos, J. C. Augusto, and D. Shapiro. Ambient intelligence—the next step for artificial intelligence. *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, 23(2):15 – 18, 2008.

S. Ray. A quick review of machine learning algorithms. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 2019.

E. Rencberoglu. Fundamental techniques of feature engineering for machine learning, 2019. URL https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114.

R. Roman, P.Najera, and J.Lopez. Securing the internet of things. *IEEE Computer*, 44:51–58, 2011.

L. Schauer, M. Werner, and P. Marcus. Estimating crowd densities and pedestrian flows using wi-fi and bluetooth. *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pages 171–177, 2014.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

S. Singh. Cousins of artificial intelligence. URL https://towardsdatascience.com/cousins-of-artificial-intelligence-dda4edc27b55.

V. Stanford. Biosignals offer potential for direct interfaces and health monitoring. *IEEE Pervasive Computing*, 3:99–103, 2004.

Y. Sun, H. Song, A. J.Jara, and R. Bie. Internet of things and big data analytics for smart and connected communities. *IEEE Access*, 4:1–1, 01 2016.

S.L. Ting and W.H. Ip. Combating the counterfeits with web portal technology. *Enterprise Information Systems*, 9:1–20, 2013.

Mark Weiser. The computer for the 21st century. *Scientific American*, pages 94–104, 1991.

J. Weppner and P. Lukowicz. Bluetooth based collaborative crowd density estimation with mobile phones. pages 193–200, 03 2013.

World Health Organization. Global status report on road safety.

L. D. Xu, W. He, and S. Li. Internet of things in industries: A survey. *IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS*, 10(4), 2014.

Y. Yuan, C. Qiu, W. Xi, and J. Zhao. Crowd density estimation using wireless sensor networks. *2011 Seventh International Conference on Mobile Ad-hoc and Sensor Networks*, 2011.

Andrea Zanella, Nicola Bui, Angelo Castellani, Lorenzo Vangelista, and Michele Zorzi. Internet of things for smart cities. *IEEE INTERNET OF THINGS JOURNAL*, 1(1), 2014.

X. Zhou. Research on wi-fi probe technology based on esp8266. In *2017 5th International Conference on Mechatronics, Materials, Chemistry and Computer Engineering (ICMMCCE 2017)*. Atlantis Press, 2017/09.