Data Collection and Analysis in Urban Scenarios

Enrico Ferrara

College of Science and Engineering
University of Derby, Derby - United Kingdom

Supervisory Team

Director of Studies	Dr. Ovidiu Bagdasar
1st Supervisor	Dr. Lee Barnby
External supervisor	Prof. Antonio Liotta Free University of Bozen-Bolzano, Italy

A submission in partial fulfillment of the requirements of the University of Derby for the award of the degree of Doctor of Philosophy

October 2021

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text. Parts of this dissertation have previously appeared in the papers listed in the primary list of publications.

Enrico Ferrara
University of Derby
2021

Acknowledgements

Firstly, I would like to thank Professor Antonio Liotta who passed on his passion for research to me and who, despite the many changes in our lives, has always been by my side.

A sincere thanks to Dr. Ovidiu Bagdasar, for his presence, availability and kindness, every contribution was fundamental to get to the end of this path.

I would like also to thank my colleagues, fellow travellers on the PhD journey with whom I shared this experience.

I am also grateful to my family and friends for their unconditional love.

Finally, I want to thank my wife Eleonora, love of my life, for always being present and supporting me every day.

Data acknowledgements

For data used in chapter 2 and 3 I would like to thank the Human Science Research Centre of University of Derby to have involved me and my colleagues in the loop of the IWUN project which was supported by the Natural Environment Research Council, ESRC, BBSRC, AHRC & Defra [NERC grant reference number NE/N013565/1].

I would like also to thank Marco Uras, Raimondo Cossu and Professor Luigi Atzori from the University of Cagliari (Italy), for collaborating on the difficult problem of wifi de-randomisation, which have led to the findings presented in chapters 4, 5 and in the appendix.

Abstract

The United Nations estimates that the world population will continue to grow, with a projection indicating a world population of up to approximately 8.5 billion people in 2030, 9.7 billion in 2050 and 10.9 billion in 2100. In addition to the phenomenon of population growth, the United Nations also estimates that in 2050 about 70% of the total world population will live in cities. These conditions increase the complexity of the services that public administrations and private companies must provide to citizens with the aim of optimising resources and increasing the level of quality of life. For an adequate design, implementation and management of these services, an extensive effort is required towards the design of effective solutions for data collection and analysis, applying Data Science and Artificial Intelligence techniques.

Several approaches were addressed during the development of this research thesis. Furthermore, different real-world use cases are introduced where the presented work was tested and validated.

The first thesis part focuses on data analysis on data collected using crowdsourcing. A real case study used for the analyses was a study conducted in Sheffield in which the goal was to understand people's interaction with green areas and their wellbeing. In this study, an app with a chatbot was used to ask questions targeted to the study and collected not only the subjective answers but also objective data like users' location. Through the analysis of this data, it was possible to extract insights that otherwise would not be easily reachable in other ways. Some limitations have arisen for less frequented areas, in fact, not enough information has been collected to have a statistical significance of the insights found. Conversely, more information than necessary was collected in the most frequented areas. For this reason, a framework that analyses the amount of information and its statistical significance in real-time has been developed. It increases the efficiency of the study and reduces intrusiveness towards the study participants. The limit that this approach presents is certainly the low sample of data that can be acquired.

In the second part of this thesis, a move on to passive data collection is done, where the user does not have to interact in any way. Any data acquired is pseudonymised upon capture so that the dictates of the privacy legislation are respected. A system is then presented that collects probe requests generated by Wi-Fi devices while scanning radio channels to detect Access Points. The system processes the collected data to extract key information on people's mobility, such as crowd density by area of interest, people flow, permanence time, return time, heat maps, origin-destination matrix and estimate of the locations of the people.

The main novelty with respect to the state of the art is related to new powerful indicators necessary for some key services of the city, such as safety management and passenger transport services, and to experimental activities carried out in real scenarios. Furthermore, a de-randomisation algorithm to solve the problem of MAC address randomisation is presented.

List of Publications

- **E. Ferrara**, A. Liotta, L. Erhan, M. Ndubuaku, D. Giusto, M. Richardson, D. Sheffield, K. McEwan, "A pilot study mapping citizens' interaction with urban nature", IEEE 16th Intl Conf on Pervasive Intelligence and Computing, (PiCom), 2018, pp. 836-841 (Aug2018).
  doi: 10.1109/DASC/PiCom/DataCom/CyberSciTec.2018.00-21.

- **E. Ferrara**, A. Liotta, M. Ndubuaku, L. Erhan, D. Giusto, M. Richardson, D. Sheffield, K. McEwan, "A demographic analysis of urban nature utilization", 10th Computer Science and Electronic Engineering (CEEC), 2018, pp. 136-141 (Sep. 2018).
  doi: 10.1109/CEEC.2018.8674206.

- L. Erhan, M. Ndubuaku, **E. Ferrara**, M. Richardson, D. Sheffield, F.J. Ferguson, P. Brindley, A. Liotta, "Analyzing objective and subjective data in social sciences: Implications for smart cities", IEEE Access 7 (2019) 19890-19906 (2019).
  doi: 10.1109/ACCESS.2019.2897217.

- **E. Ferrara**, L. Fragale, G. Fortino, W. Song, C. Perra, M. Di Mauro, A. Liotta, "An AI approach to collecting and analyzing human interactions with urban environments", IEEE Access 7 (2019) 141476-141486 (2019).
  doi: 10.1109/ACCESS.2019.2943845.

- M. Uras, R. Cossu, **E. Ferrara**, A. Liotta, L. Atzori, "Pma: a real-world system for people mobility monitoring and analysis based on WiFi probes", Journal of Cleaner Production, 2020, 0959-6526 (2020).

doi: 10.1016/j.jclepro.2020.122084.

- M. Uras, R. Cossu, **E. Ferrara**, O. Bagdasar, A. Liotta and L. Atzori, "WiFi Probes sniffing: an Artificial Intelligence based approach for MAC addresses de-randomization", IEEE 25th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), Pisa, Italy, (2020).
  doi: 10.1109/CAMAD50429.2020.9209257.

- **E. Ferrara**, M. Uras, L. Atzori, O. Bagdasar and A. Liotta, "Mobility Analysis during the 2020 Pandemic in a Touristic city: the Case of Cagliari", 2021 IEEE IoT Vertical and Topical Summit for Tourism (IoT-VTST'21), September 2021.

Contents

Acknowledgements **iv**

Abstract **viii**

List of Publications **ix**

List of Figures **xvii**

List of Algorithms **xix**

List of Tables **xxi**

1 Introduction **1**

1.1 General Context 1

1.2 Research Motivation 3

1.3 Research Aims and Objectives 6

1.4 Major Contributions of this Research Study 7

1.5 Thesis Outline 8

2 Mapping Citizens' Interaction with Urban Areas **11**

2.1 Introduction 12

2.2 Related Work 15

2.2.1 Data challenges in social science studies 15

2.2.2 Mining Objective and Subjective data 15

2.2.3 App-based studies on the connection wellbeing-nature . . . 16

2.3 Methods 17

2.3.1 Shmapped 17
2.3.2 Data collection 18
2.3.3 Data cleaning and pre-processing 20
2.3.4 Text analysis 22
2.3.5 Image analysis 24
2.3.6 Time Analysis 25
2.4 Dataset characterisation 26
2.4.1 Participants' description 26
2.4.2 Participants' interaction with Shmapped 26
2.4.3 Participants' wellbeing 29
2.5 Features noticed by the users 29
2.5.1 What do the images say? 29
2.5.2 What does the text say? 31
2.6 Time spent in green spaces 37
2.6.1 Average time spent in green spaces 37
2.6.2 Age and gender distribution in park utilisation 39
2.7 Comparison between objective and subjective interaction 40
2.8 Summary of key results 41

3 A framework for improving subjective data collection 45

3.1 Introduction 46
3.2 Related work 48
3.2.1 Chatbots 49
3.2.2 Analytic Frameworks 50
3.2.3 Complexity of data collection 51
3.2.4 Objective and Subjective Data 51
3.3 Proposed Framework 52
3.3.1 Overview 52
3.3.2 Client side 53
3.3.3 Server side 56
3.3.4 Data privacy and data security 58
3.4 Internal and external contexts analysis 59

3.5 Evaluation through a case study 61
3.5.1 Evaluation method 61
3.5.2 The IWUN Dataset 62
3.5.3 Setting threshold goals 62
3.5.4 Setting statistical goals 65
3.6 Summary of key results 66

4 A passive data collection framework exploiting Wi-Fi protocol 69

4.1 Introduction 70
4.2 Related works 72
4.2.1 Localisation Techniques 73
4.2.2 Trajectory Tracking 73
4.2.3 Crowd Density and Flow 74
4.3 The PmA system 75
4.3.1 PmA Stations 76
4.3.2 PmA Platform 79
4.3.3 Privacy aspects 80
4.4 Data Analysis 82
4.4.1 Counting 85
4.4.2 Site Returns 85
4.4.3 Site Permanence 86
4.4.4 Origin/Destination Matrix 87
4.4.5 Crowd Density 88
4.5 Experimental analysis 92
4.5.1 Experiments setup 94
4.5.2 Experimental results 96
4.6 Summary of key results 103

5 Mobility tracing through data analysis 105

5.1 Introduction 105
5.2 Related work 107
5.3 Data Science Pipeline 109

5.3.1 Data acquisition 109
5.3.2 Anomaly detection 109
5.3.3 Data selection 110
5.3.4 Data Modelling 110
5.4 Data Analysis 111
5.4.1 City Traffic Analysis 111
5.4.2 Touristic Flows Analysis 116
5.5 Summary of key results 116

6 Discussion and Conclusion 119
6.1 Summary of results 119
6.2 Future work 122

A A de-randomisation algorithm for Wi-Fi data collection 125
A.1 Introduction 125
A.2 De-randomisation overview 127
A.3 Background 129
A.3.1 MAC address structure 129
A.3.2 MAC randomisation 131
A.3.3 Related past works 132
A.4 The sniffer 136
A.5 The proposed de-randomisation algorithm 139
A.5.1 Features extraction 142
A.5.2 Pseudo-random MAC filtering 147
A.5.3 Frame clustering 148
A.6 Results in a controlled environment 151
A.7 Results on-board of city buses 152
A.7.1 Data acquisition setup 153
A.7.2 Dataset analysis 154
A.7.3 Performance analysis 156
A.8 Conclusions 159

Bibliography 159

List of Figures

1.1 Research work and thesis flow. 9

2.1 Preview of the extracted geofences. 19
2.2 The study area for time analysis. 22
2.3 The smallest green spaces taken into account in the study. 25
2.4 Age distribution of the sample dataset. 27
2.5 Participants' companionship / social interactions. 27
2.6 Participants' feelings while interacting with their surroundings. . . 28
2.7 Density of the users' feelings and the associated grades. 28
2.8 Top 10 labels for each category of images. 30
2.9 Users' textual observations clustering. 33
2.10 Classification of the textual observations into the themes. 34
2.11 Age classification of textual observations. 35
2.12 Gender classification of textual observations. 35
2.13 Endcliffe Park people interaction. 38
2.14 Age groups interactions with top 10 most visited green spaces. . . 39
2.15 Gender groups interactions with top 10 most visited green spaces. 40
2.16 Age groups objective and subjective interactions. 42
2.17 Gender groups objective and subjective interactions. 43

3.1 Interaction diagram between clients, chatbots, and server. 54
3.2 Chatbot Interface. 55
3.3 Analysis flow. 58
3.4 Mapping functions. 60
3.5 Context analysis and representation. 60

3.6 Comparative results when setting threshold goals. 64
3.7 Information gain evolution over time. 64
3.8 Distribution of user's feedback. 65
3.9 Comparative results when setting statistical goals. 67

4.1 Macro-architecture of PmA system. 76
4.2 Detailed system architecture. 77
4.3 Outdoor PmA station used in the pilot study. 77
4.4 Storage architecture. 79
4.5 MAC address detail. 82
4.6 Data Flow. 83
4.7 Target position estimated by RSSI-based algorithm. 90
4.8 Target position estimated by TDOA-based algorithm. 93
4.9 Trajectory comparison using the RSSI-based algorithm. 93
4.10 Map of the controlled scenario area. 94
4.11 PmA Stations used for tests in the controlled scenario. 94
4.12 O/D Matrices in different time bands. 98
4.13 Number of returning devices (Turin). 99
4.14 Crowd Density in the truffle fair in Alba. 99
4.15 Uniquest MACs Distribution in University of Cagliari. 100
4.16 Number of returning devices at University of Cagliari. 101
4.17 Number of stationary devices at University of Cagliari. 102
4.18 Crowd Density in University of Cagliari. 102

5.1 Traffic measurement stations in the city. 106
5.2 Data pipeline; from raw data to a clean dataset. 109
5.3 Inductive loops of traffic sensors installed in Cagliari. 110
5.4 Comparison between 2020 and 2016-2019 average traffic volumes. 111
5.5 Change in traffic volumes in 2020. 112
5.6 Weekly volume traffic distribution. 114
5.7 Daily traffic distribution for each of the eight pandemic periods. . 115
5.8 2020 traffic stations volumes and touristic flow. 117

A.1 Randomisation implementation timeline. 126
A.2 MAC address structure. 130
A.3 Probe requests sensor with multiple interfaces. 137
A.4 Comparison of captured frames number with different configurations. 138
A.5 Sequence of frames generated by two devices. 141
A.6 De-randomisation algorithm workflow. 143
A.7 25 minutes Probe Requests acquisition. 146
A.8 Identification of Pseudo Virtual frames: ROC curve. 148
A.9 Semi-anechoic chamber tests: number of devices. 151
A.10 Setup of the anechoic chamber tests. 152
A.11 Power distribution. 154
A.12 Vendors distribution. 155
A.13 Bus acquisition and data cleaning problem. 156
A.14 Number of devices estimated for each dataset in the bus scenario. 157
A.15 Device traces with a pool of used MAC addresses. 158

List of Algorithms

1 External Context Evaluation. 61
2 Site Returns Algorithm. 86
3 Site Permanence Algorithm. 87
4 PmA Derive Positions algorithm. 89

List of Tables

2.1 Labels from training data. 23

2.2 Example of labelling for an image. 24

2.3 Gender distribution. 26

2.4 Number of labels for participant categories. 30

2.5 Text clustering. 32

2.6 Average time spent in parks by user. 36

2.7 Average time spent inside green spaces by park. 36

4.1 Recent literature for Wi-Fi probe analytics. 72

4.2 High-gain antenna details. 78

4.3 Error evaluation in meters. 96

5.1 Period details. 112

A.1 OS behaviour comparison for MAC address randomisation. 131

A.2 Used notation. 142

A.3 Example of information extracted from the Probe Requests. . . . 144

A.4 Features extracted from Probe Request packets as sparse matrix. . 144

A.5 Presence rate of the IEs on the frame total number. 145

A.6 Test devices used in the semi-anechoic chamber test. 152

A.7 Details of the Datasets used. 153

A.8 Estimation results errors. 158

Chapter 1

Introduction

This chapter introduces the thesis contents. In the first part, the general context and the motivation of the study are presented, then the research aims and objectives are stated to highlight the main contributions of this work. This chapter ends with a brief thesis outline and a summary of the contents of each chapter.

1.1 General Context

As highlighted in numerous United Nations reports [1], world population will continue to grow rapidly, with estimates of about 8.5 billion in 2030, 9.7 billion in 2050 to reach 10.9 billion in 2100. At the same time, the urban population (percentage of people living in urban areas) in Europe will increase from today's estimate of 75% to 80% in 2050 [2]. This process is not only present in Europe, the United Nations estimates that by 2050 about 68% of the world population will live in cities [3]. Clearly, with more population concentrated around urban areas, the complexity of managing cities will increase proportionally. It is therefore, necessary to develop tools that make it possible to improve the quality of services offered to citizens by public administrations and private companies. In addition, such tools would also allow the efficient use of resources "to achieve a better and more sustainable future for all", which is one of the key sustainable development goals defined by the United Nations.

New technologies will enable a transformation into real smart settlements, from the smallest and rural cities to the largest and most organised ones. In fact, the European Commission defines them as "cities that using technological solutions improve the management and efficiency of the urban environment"[4]. Six different aspects of a smart city that need to be developed have been identified:

Smart People: people (citizens) must be involved, bottom-up (from bottom to top) decision-making and participatory policy need to be integrated.

Smart Governance: the administration must give priority to human capital, environmental resources, relationships and community assets.

Smart Economy: the economy and urban commerce must be aimed at increasing productivity and employment within the city through technological innovation. An economy based on participation and collaboration and which focuses on research and innovation.

Smart Living: the level of comfort and wellbeing that must be guaranteed to citizens linked to aspects such as health, education, safety, culture, etc., are also of top importance.

Smart Mobility: intelligent mobility solutions, from e-mobility to sharing mobility to other forms of mobility management, must look at how to reduce costs, reduce environmental impact and optimise energy use.

Smart Environment: sustainable development, low environmental impact and energy efficiency are priority aspects of the city of the future.

Each of these aspects requires abundant and reliable data on which to make decisions, in order to make informed decisions to achieve the desired outcome. This thesis presents novel ways of acquiring and analysing data employing cutting edge machine learning and data science techniques. These methods aim to maximise the insight gained from studies conducted in urban areas. The contributions of this thesis are detailed in the following subsections while shortcomings and research gaps in previous works are indicated specifically in each chapter.

1.2 Research Motivation

The collection of data showing the interaction of people with the urban environment is a complex task, with many existing studies showing the limitations [5]. For example, there are privacy-related restrictions on the data acquisition by crowds in public spaces, the difficulty of building temporary experimental setups that realistically represent real-life scenarios, and ethical constraints regarding the creation of stressful experimental environments. In this work different technical solutions are presented, most of them customised to the application context.

The first part of the work is a social study in collaboration with the Human Research Centre, where data on the interaction of citizens with green areas in a city had to be collected and analysed, in order to obtain insights regarding the impact of this type of interaction on citizens' wellbeing. The data was collected through an ad-hoc smartphone app where the study participants interacted and sent their comments and/or pictures through a chatbot. The app made it possible to collect not only the subjective data uploaded by users but also the objective data extracted from the phone's sensors. Having these two categories of data collected, data fusion helped us obtain different levels of analysis and comparison.

Before the analysis and the extraction of insights, it was necessary to find ways to address incomplete data, lack of data, or incorrect data which could affect the statistical significance of the study. This step enabled us to learn how to perform large-scale social studies and to develop useful analysis techniques. The results obtained from subjective and objective data include certain behaviour patterns of citizens with their surroundings. Some limitations of this approach have also become evident. The questions sent by the chatbot were the same each time and sent indistinctly to some random people who entered inside the areas under examination. This methodology led us to collect data with a low ratio of answers number over insights obtained. An example is that the data about popular areas of the investigation was more than that required to find insights, while for less frequented areas insufficient data was collected. That means the methodology was intrusive in all the cases where the insight threshold was reached and ineffective for areas where the threshold was not achieved.

The limitations of the first study have shaped and informed the second part of the research work in this thesis. The goal was to submit the questions more precisely, in order to increase the statistical significance of the collected answers and minimise the intrusiveness created by the chatbot querying the users. The whole system was then redesigned, using a more intelligent and reactive approach in selecting the questions to be sent to users. The first step was to design a new app using artificial intelligence (AI) so that the chatbot can change the order of the questions asked in order to maximise knowledge gain. However, the chatbot's AI does not have a complete view of the system and of the responses collected globally, thus it has to interact with the server where the data is analysed on arrival and where the study's global results are updated in real-time.

Depending on the results collected and the relative statistical significance, the system decides which questions should be sent and to which users, giving instructions to the chatbot which, through its AI, decides which questions to submit to the user. The result is an interactive data analysis framework for urban environments. This system is then tested on the previous case study, creating a simulator that reproduces the arrival of information to the server in the same way as it happened with the static system previously used. The differences between the two approaches will then be analysed, and the significant improvements of the solution that uses AI in data analysis will be appreciated. At this point another important research question arose:

Is it possible to obtain similar insights without the need to involve directly people and thus obtain data from all the people that interact within a certain area instead that only from a sample of them?

A paradigm shift was therefore made, passing from active to passive data acquisition. The active approach has in fact the main disadvantage that users must play an active role in data collection because they must at least install the app and, in some cases, also provide the requested input. Furthermore, this approach often requires awarding prizes to the users involved in order to obtain minimum user samples. The passive approach, on the other hand, does not require users to actively participate and is, therefore, less intrusive and the data collected takes into account almost all the people present in the under-examined areas. Over the past 15 years, a significant research effort has been made on locating people using sniffing of packets sent by devices using Wi-Fi technologies. This system plays an important role thanks to its low implementation costs but nevertheless, this research field requires significant efforts in order to obtain practical, robust and accurate solutions. In particular, one needs to devise adequate processing that, starting from the raw data, can generate the information required to address the challenges of the city. Furthermore, the data collection form should respect the privacy of the persons being monitored.

In the second part of this thesis, the design, development and test of the data acquisition and analysis system were carried out by sniffing the packets sent by the Wi-Fi interfaces of smartphones. Respecting privacy and all GDPR [6] requirements, this data enables tracking and locating people, deducing key information on crowd mobility. Exploiting the weakness of the Wi-Fi protocol allowed the extraction of insights and data from each device located in the monitored area. A global architecture is necessary, from the acquisition, where is important to develop specific sensors able to detect the different messages sent from devices to the access point and vice-versa, to the data analysis where a specific pipeline cleans the data and makes complex computations in order to discriminate between the different devices.

An entire chapter is dedicated to the work done in the design and implementation of this framework while in the appendix there is an extract of the journal paper sent, and currently under review, relating to the work necessary to solve the problem of randomisation of the MAC address. The de-randomisation algorithm extends the work reported in this thesis bringing it to be fully exploitable. In fact, it allows, within the area under examination, to acquire and analyse all messages sent through the Wi-Fi protocol even if the MAC address randomisation is used. In general, despite the remaining limitations, this approach is very effective in extracting insights on urban mobility.

The next interest point addressed concerns the integration of different data sources which is still necessary to obtain more objective and above all more complete insights, as it is not always possible to identify different phenomena with a single data source. In the last part of this thesis work, a pilot study is presented where a large dataset of mobility traces is analysed, generated by 98 traffic sensors scattered around the city, and operational since 2016 which have been available from the municipality of Cagliari as open data on their website. In this part of the research work, the cleaning and data processing methodology is not simply introduced, but the impact that the Covid-19 pandemic has had on urban mobility during the year 2020 is also reported.

The results obtained are rich and while some were intuitively predictable (drastic reduction in traffic volumes during the quarantine), some interesting insights have emerged. For example, it has been noted that following an initial traffic reduction of 76% at the first lockdown (March 2020), subsequent restrictions have led to less drastic changes. Also, while the absolute traffic volumes have roughly followed the evolution of the pandemic, the weekly traffic patterns have drastically changed over time, while the daily ones have maintained greater consistency. The traffic traces were also compared with the official tourist presence data, which made it possible to identify the traffic stations most affected by the mobility of tourists.

1.3 Research Aims and Objectives

This thesis aims to study how data science techniques can improve data collection and analysis in urban contexts. This topic leads to the following research goals.

- To propose a novel way of analysing data from social sciences studies involving huge numbers of people.
- To improve the collection of statistically significant subjective data and reduce intrusiveness in active data collections via smartphone apps.
- To design a passive analysis data collection framework using the Wi-Fi protocol.

- To propose a custom pipeline to clean up the data collected by traffic sensors and to extract information that could be used as another source of information to be merged into a global data acquisition framework.

- To validate all previous points in case studies where data were collected from real-life scenarios.

- To propose a de-randomisation algorithm to identify the probes sent from the same device even if the MAC address change randomly every few packets solving the present limitation in passive data collection through Wi-Fi.

1.4 Major Contributions of this Research Study

1. *A novel way of analysing data from social sciences studies involving huge numbers of people is proposed.* Exploiting the data collected through a smartphone app, the objective data coming from the smartphone sensors and the subjective data coming from users input are fused in order to extract insight about the social study. The resulting work was published in two conferences [7, 8] and at "IEEE Access Journal"[9].

2. *In active data collections via smartphone apps, the collection of statistically significant subjective data was improved and the intrusiveness was reduced.* Using artificial intelligence (AI), objective and subjective data are analysed in real time to examine the subjective data test matrix. Through this analysis, it is possible to determine which action maximises the information gain. For example by avoiding submitting questions for which enough answers have already been collected to be sufficiently meaningful and asking for others for which there are not yet enough. The resulting work was published at"IEEE Access Journal"[10].

3. *A novel analysis is proposed in passive data collection exploiting the Wi-Fi protocol.* The MAC address sent by the devices in the probe request during the active scan for the discovery of the access point has been collected. Several new key insights into people's mobility were extracted from that

data such as crowd density by area of interest, the flow of people, residence time, return time, heat maps, origin-destination matrices and estimation of the location of people. The resulting work was published at "Journal of Cleaner Production"[11].

4. *A custom pipeline is proposed to clean up the data collected by traffic sensors and extract information from it.* Taking into account a large dataset of mobility traces by open data, a significant effort was required to pre-process the raw data, which otherwise would not be directly used due to problems arising from the data collection and transmission process. Then a different perspective is shown, focusing on traffic volume, patterns and information relating to tourists. The resulting work is accepted at "2021 IEEE IoT Vertical and Topical Summit for Tourism"[12].

5. *To address the randomisation problem, a de-randomisation algorithm is proposed to derive which probes are sent by the same device even if the MAC address change randomly every few packets.* In addition to the MAC address, the values relating to the different information elements sent are also collected within the probe request. With a complex analysis of these fields, it is possible to identify a fingerprint for each device and understand its behaviour over time. The resulting work was published at "IEEE 25th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)"[13].

1.5 Thesis Outline

In Figure 1.1 is shown how this research thesis work is structured and the relative chapters where each part is described deeply.

Chapter 2 introduces how it is possible to apply data science techniques in social studies, acquiring objective and subjective data from an app to derive insight useful to improve the quality of city services and people wellbeing.

Chapter 3 introduces the work done to improve the statistical significance in the

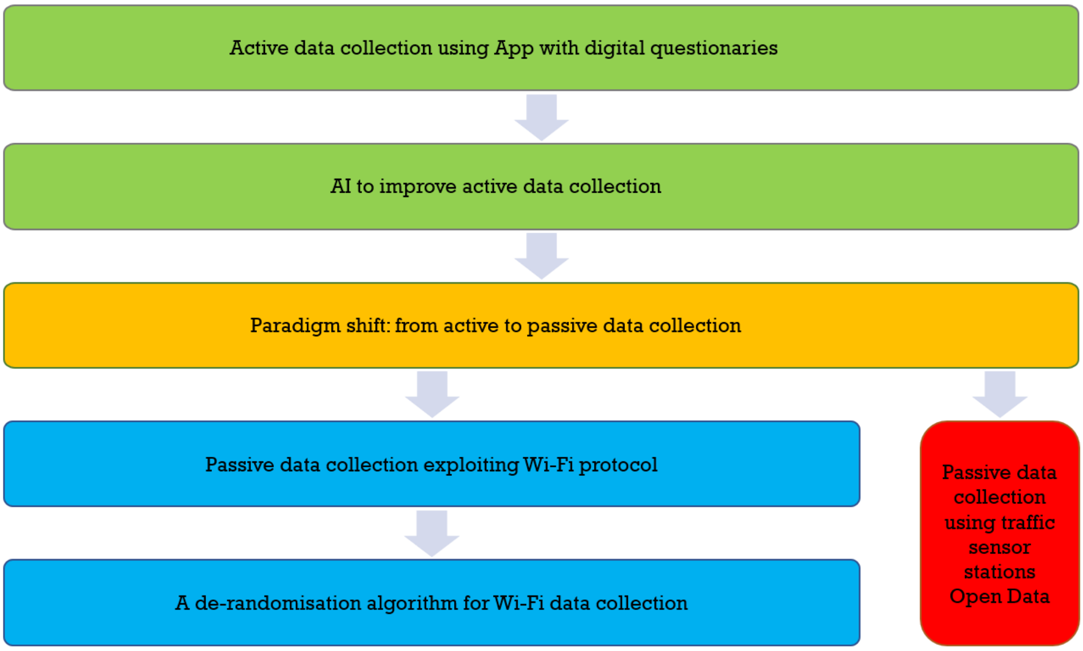

Active data collection using App with digital questionaries
AI to improve active data collection
Paradigm shift: from active to passive data collection
Passive data collection exploiting Wi-Fi protocol
A de-randomisation algorithm for Wi-Fi data collection
Passive data collection using traffic sensor stations Open Data

Figure 1.1: Research work and thesis flow.

data acquired using the active way through an app.

In Chapter 4 a different approach is presented, which exploits the Wi-Fi protocol to acquire more data in a passive way with less intrusiveness.

In Chapter 5 the pipeline to clean and analyse the traffic data coming from Open Data is shown.

Finally, Chapter 6 concludes the work by discussing the potential future directions for the work presented in this thesis.

In addition, in appendix A the algorithm to solve the problem of MAC addresses randomisation is presented.

Chapter 2

Mapping Citizens' Interaction with Urban Areas

The advances in the Internet of Things and crowd-sensing enabled the collection of vast amounts of urban data, allowing researchers to better understand how citizens interact with cities and, in turn, improve human wellbeing in urban environments. In vast urban areas, collecting statistically significant data is a daunting task: new data-collection methods are needed, along with processes for fusing objective (heterogeneous) data (*e.g.* people location trails and sensors data) with subjective (perceptual) data (*e.g.* the citizens' quality of experience collected through feedback forms). In collaboration with the Human Research Centre of University of Derby, as part of the IWUN project[14], in this chapter is presented a pilot study carried out in Sheffield (U.K.) which aims a better understanding of the interactions between citizens and urban green spaces.

With the help of a smartphone app, both objective and subjective data were collected. Location tracking was recorded as people entered any of the publicly accessible green spaces. This was complemented by textual and photographic information that users could insert spontaneously or when prompted (when entering a green space). By employing data science and machine learning techniques, the main features observed by the citizens through both text and images were identified. In addition, the time spent by people in the parks and the areas of greatest interaction was also analysed. This chapter allows us to gain an overview of certain patterns and the behaviour of the citizens within their surroundings and it proves the capabilities of integrating technology into large-scale social studies.

Contributions The material presented in this chapter is related to joint work involving the Data Science Research Centre (in the people of Laura Erhan, Maryleen Ndubuaku, Enrico Ferrara and Antonio Liotta who was the centre coordinator), the Human Sciences Research Centre (in the person of Miles Richardson, David Sheffield, Fiona J. Ferguson) and the Department of Landscape Architecture represented by Paul Brindley (PB).

The Human Sciences Research Centre, which was directly involved in the IWUN project, has contributed to the work in several ways. First, they conceptualised the app used in the study, then implemented by the developers of Furthermore Ltd. Secondly, they evaluated the social and psychological aspects of the results obtained through the analyses carried out by the Data Science Research Centre. PB, also directly involved in the IWUN project, was of fundamental importance to correctly interpret the starting data and the functioning of the app.

As for the analyses carried out by the Data Science Research Centre, the work in its entirety was carried out as a team. The individual analyses were instead developed separately, writing the code necessary for the specific analysis, to then be merged again and analysed in teams for global insights. Specifically, the author of this thesis dealt with the data pre-processing and the space-temporal analyses.

2.1 Introduction

Human interaction with cyber-physical systems [15] is an important issue which, thanks to the Internet of Things and the consequent digitisation of the physical world, have prompted researchers to carry out several multidisciplinary studies. A paradigm centred on the social side of the IoT [16] emerges thanks to the desire to harmonise the interaction between society and smart objects.

Although operating with an increasingly stringent administrative budget is a primary constraint, the main goal remains to provide a better use of public infrastructures and a better quality of services to the citizen with consequent improvement in the quality of life [17, 18]. Thanks to these key interventions it is possible to directly influence urban health and wellbeing and this chapter presents how data science and machine learning techniques can be used to improve social studies. In the past, this type of study was almost always based on data collected manually through paper diaries or self-administered forms [19].

Processing the resulting data was complex, required long processing times and took into account only the subjective data collected. In one of these studies [20], using statistical techniques, S. Ruiz-Correa et al. analyse the perception of young people of a developing country, relative to the city in which they live. Instead, the presented work uses data science tools to discover patterns and create correlations that may not be easily identified with traditional statistical tools.

In addition, objective data is also collected, using the Shmapped smartphone app created ad hoc in the context of the IWUN project, where the aim was to monitor the interaction of citizens with green spaces within the urban context.

The study involved 1,870 people around the Sheffield area (UK), where 760 green spaces of interest to the study were identified. For 7 and 30 days (2 different versions of the study) the people involved used the application that collected both objective data such as position and type of activity carried out, and subjective data, as thanks to the presence of a chatbot, all the user was asked to indicate various parameters (wellbeing, personal feelings, type of social interaction, users' observations on the surrounding environment). The user could also upload texts and images freely throughout the day. This app, therefore, works not only as a data acquisition, but also as an intervention tool, as it pushes the study participants to notice, photograph and describe what they saw in the surroundings.

According to some studies [21, 22] being in contact with nature and noticing the details can improve people's wellbeing. Therefore, if on the one hand it is possible to improve the wellbeing of people, stimulating them to contact with nature itself, on the other hand it is possible to collect their interactions and sensations, and by using suitable frameworks it is possible to manage cities in a more intelligent way, improving the quality of life [23, 24].

The biggest challenges with such frameworks are the complex processes involved in planning, collecting and analysing large amounts of data. These processes can then be improved by using a large-scale IoT infrastructure [25]. The work presented therefore exploits the personal smartphone as an IoT object and in which through the app it is possible to collect objective data (sensor information) and subjective data (user interactions), thus improving the traditional collection of data in social science studies. In fact, by merging these data, it is possible to discover different ways of interacting by the citizen with the urban environment.

One of the substantial differences from traditional analysis is that it is possible to monitor at the moment of interaction, collecting both subjective and objective information on the specific moment. For each subjective information collected, specific objective information from the sensors are associated and are used to define the interaction times within the urban space under analysis, the speed of movement and other parameters. By this approach it is possible to automate the collection and analysis of data, increasing the size of the study and the collection of information from more people, magnifying the sample subject to the study.

The system presented includes an initial phase of collecting and pre-processing the data generated by users, subsequently these data are processed, analysing the observations and photos sent by users as well as making a careful analysis on the paths and times of spent inside the study's areas of interest.

Part of the investigation was to map observations into topic groups against previous research topic categories to notice good things in nature [26]. With regard to the traces created by the localisation points of users, the time spent by the different areas is calculated and compared with the information extracted from the analysis of subjective data. This leads to effective data fusing, which enables added value on the information relating to the interaction of users with the surrounding environment. Having this information, it is therefore possible to take action to improve services and quality of life for citizens more effectively.

In Section 2.2 an overview of the related work is provided; Section 2.3 describes the methodology; Section 2.4 characterises the dataset used; Section 2.5 outlines features noticed by users; Section 2.6 examines the time spent by users in green spaces; in Section 2.7 an analysis of the use of the park based on gender and age is presented, and the key results of the study are reported in the Section 2.8.

2.2 Related Work

2.2.1 Data challenges in social science studies

Most definitions and studies of Big Data in cities are limited by the volume attribute of Big Data. It has become a trite definition that anything which does not fit into an Excel spreadsheet or cannot be stored in a single machine is Big Data [27]. For instance, the study in [28] analysed half a million waste fractions to identify inefficiencies in waste collection routes. In [29], Anantharam et al. analysed social textual streams comprising 8 million tweets to extract traffic events in the city of San Francisco Bay Area. Though this study may not fit the volume bracket based on the number of participants of the study, it copes with other inherent characteristics of Big Data which make it challenging such as its variety (composing of structured and unstructured data), exhaustivity (attempts to capture all the population), scalability (can rapidly expand in size), and relationality (has common fields that can be correlated) and messiness [30].

2.2.2 Mining Objective and Subjective data

Acquiring data remains one of the most demanding and complex tasks, as there is a myriad of possibilities both for the acquisition method and for the data itself chosen. In fact, objective data can be obtained when IoT devices are interrogated that generally report physical phenomena in the digital world while subjective data must, in any case, be acquired by people using the different systems.

The choice of which devices and data to obtain is clearly influenced by the type of study to be carried out; for example, in [31] Fujiki et al. collected accelerometer data to evaluate the user's metabolic activity. A study more focused on smart cities, in which urban mobility was monitored in real-time is certainly the study by Calabrese et al. [32] where GPS position data of buses, taxis and pedestrians were collected. The data collected through sensors however suffer from data quality problems such as uncertainty (sensor accuracy, missing readings), inconsistency and redundancy in the data [33]. They also do not take into account the human component, understood as a given subject by human beings in interaction with smart objects present within smart cities [34].

With the massive spread of social networks, it is easier to collect subjective data such as, for example, tweets relating to certain [35] events. It is therefore possible to collect specific subjective data for the interest of the research field, limiting the volume of data collected but improving the analyses and making them richer, more diversified and complementary for smart cities [36].

In this chapter, a hybrid approach is used in data collection: objective data from GPS and sensors is gathered, and at the same time, subjective data such as textual information and images are collected, entered via the app by the participants. Similar approaches have been used by MacKerron et al. in [37] where participants were asked to indicate their wellbeing levels at random times over a 24 hours period, and at the same time the data relating to the users' location was also acquired. Here, also the integration of the text analysis relating to the observations on the environment, the analysis of the images uploaded by the user, and of elements related to the time spent and position in a certain area are done.

2.2.3 App-based studies on the connection wellbeing-nature

With IoT and smartphones, it is now possible to automatically collect large samples of both subjective and objective data. This is more cost-effective and involves larger datasets and, in turn, yields more statistically robust findings [38].

Mappiness [37] is a social App designed as an intervention tool to enhance happiness as an element of wellbeing. In Mappiness, participants are asked to report their wellbeing at random times during the day, whilst having their location tracked. Urban Mind [39] is another social App, designed to examine how exposure to green spaces impacts mental wellbeing in real-time. In that study, there were seven prompts a day for assessing wellbeing in urban areas. The set of questions the users had to answer was dependent on their location (indoors/outdoors).

In both cases, most of the collected data was done when the participants were indoors, as they only spent at most 14 percent of their time outdoors, making it challenging to collect the data in green spaces, where the reported level of happiness is actually higher. This was seen as a major limitation in these two Apps. In an attempt to address this issue and optimise Shmapped for data collection, the green spaces were structured into geofences and the users were prompted to observe their environment upon entering one. Thus, the reliability of the study concerning the interaction with nature is improved as people are prompted to give details of their experience when in green spaces.

In general, it is possible to state that previous research work was done focusing only on objective or subjective data. In cases where both were considered, the observation period and the monitored geographic area were short. The work presented here wants to fill this research gap, applying the analyzes both on objective and subjective data, over a period and in an extended test area.

2.3 Methods

2.3.1 Shmapped

In order to acquire data from people involved in the study, the Human Science Research Centre conceptualised an app then implemented by the developers of Furthermore Ltd. Using that app, all the data were acquired and lately analysed as explained in this chapter with the aim to understand how people interact with green areas within the urban context. The app name is "Shmapped" which is the contraction of "Sheffield Mapped" as the study was located in the urban area of the city of Sheffield. Shmapped interacts in a friendly and engaging way with users, through a chatbot and it is used as a dual utility tool.

The app is used as an objective data collection tool, gathering the position when the user enters the green areas of interest, identified through geofences. Using data from device sensors, one can distinguish the type of activity and classify it with different labels (stationary, running, walking, etc.). Along with objective data, the app through the chatbot also collects subjective data through textual observations and images uploaded independently by the user. In addition, it works as an intervention tool by pushing users located inside the geofences to notice the surrounding environment, through questions asked by the chatbot.

In a first version, the study lasted 30 days but having seen a drastic decrease in user interaction with the app, the trial period was reduced to 7 days, in which the participant was asked to interact with the app. spontaneously or at least to answer the questions the chatbot asked. One of the most relevant data in any interaction is the level of wellbeing, which wants to be monitored to understand how much the interaction with urban green areas affects daily wellbeing.

To monitor and measure both wellbeing and the connection with nature, three different questionnaires are administered (again digitally through the app). The first is administered at the start of the study, where demographic data and the assessment of individual differences and wellbeing are checked; another after 7 days upon completion of the study, and a third at follow-up (1 month for the 7-day version or 3 months for the 30-day version). Through these questionnaires, one can understand the wellbeing variation of participants in this study.

2.3.2 Data collection

There are two main types of data gathered from users: Subjective Data and Objective Data. GPS locations of participants were tracked within digital geofences (circular areas comprising the green spaces of interest as shown in figure 2.1), with data then being recorded once participants entered the more detailed geography of publicly accessible green spaces (provided by Sheffield City Council). The use of the geofences allowed phones to be woken from standby alongside more accurate GPS recording. Specific information about the data collected are:

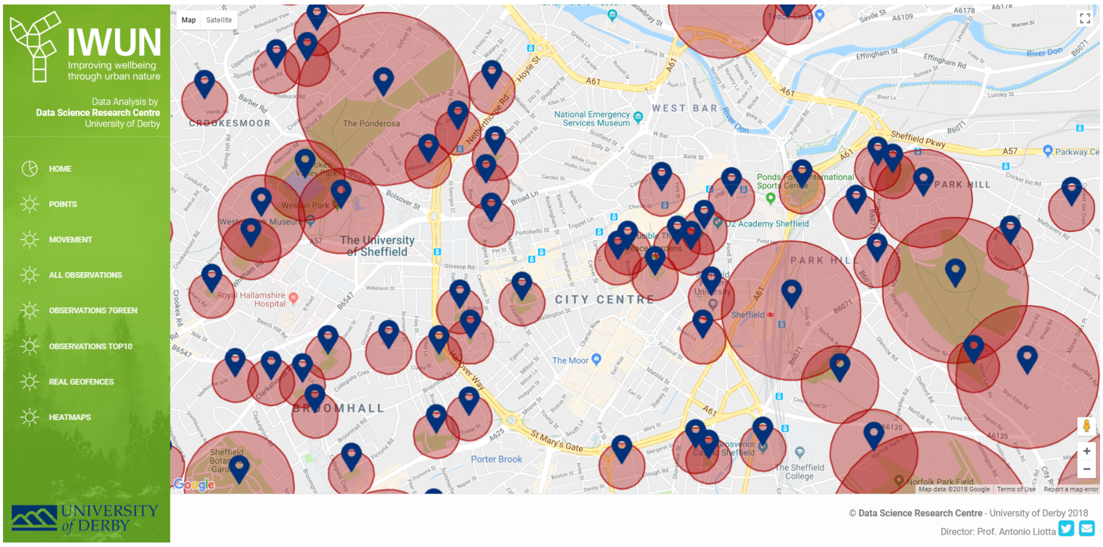


Figure 2.1: Preview of the extracted geofences.

- Locations (Objective Data)
  - GPS Data. Location and speed data is used to infer users' dwelling time in green spaces.
  - Derived Data. The information provided by sensors is used by the App to evaluate user activity, like is_moving and activity_type (classified into four main categories: still, on foot, in-vehicle and unknown).
- Observations (Subjective Data)

  During the study, the App asks users to mention "good things they noticed" around them. When inserting a comment, the App asks for additional information in order to assess their experience. Data collected includes:
  - Comment about what they noticed;
  - Picture (optional);
  - Why they are in that place ('whyThere');
  - With whom they spend time ('whomWith');
  - How built-up the place is ('howMuch');
  - How they feel in the moment ('howFelt').

2.3.3 Data cleaning and pre-processing

One of the first steps done at the beginning of the interaction with the dataset is data cleaning. Parts of the collected data were irrelevant for the study and problem at hand. For example, there were users who registered and took part in the study but were not living in Sheffield, UK. As the study was focused on this specific city, their data had to be filtered out. The subjective user responses included free text, images, or a mix of both, but also controlled input such as: whom they were with, how they felt, why they went there, how built-up the environment was. The information were fused through a mix of semantic text and image analysis as well as correlating the whom, why and how. The objective data includes mainly the location points and other sensor information, which were used as the starting point to infer things like dwelling time and type of activity. The types of cleaning or filtering which were undertaken are shortly described below based on the category of data they belong to.

Users

For the study, participants were split into two categories: green (70%) and built (30%). The former group was prompted to notice good things about nature. The latter group recorded their observations regarding the built environment, a condition which was included by the psychology researchers as a control group. For this analysis, parts of the data were split according to the built/green criteria, with an emphasis on the green. This is due to the focus on gaining insight into the citizens' interaction with the natural surroundings. The total number of registered users in the App was 1870. Out of these, 580 were part of the built group and 1290 of the green group (69%). It is important to note that the numbers of unique users in the different types of analysis turned out to be lower. This is because not all people who registered went on to use the App or provide data. Furthermore, some of the users who signed up were not living in Sheffield. They were filtered out by using the postal code provided at registration.

Observations

Observations are referred to the text comments and the images taken by the users. It is important to note that only 418 entries out of 5626 had a timestamp associated, meaning that they were recorded at the time of the observation (when the user was prompted to notice the surroundings and input the data). The rest were entries made later during the day, mainly in the evening after the reminder given by the App. The App asked the users to manually input their location, but in most cases, the field was left blank. Some of the earlier analysis conducted was focused on these 418 entries; *i.e.* the parks with the most registered observations. No optimal way was found to reconnect the rest of the observations with their location. Possible ways of achieving this could be: looking at the comments to check if a location is mentioned and see what parks were visited by the users during the day. Problems arising are required knowledge on the park names and the possible variations, multiple parks visited in one day, as well as general comments or information which cannot be tied to one specific area.

For the textual analysis, the data provided by the green group were analysed and it is 4226 entries from 718 users. Location was not taken into account here as the focus was rather on classification and feature extraction. To have a better clustering performance, the text to only include the green users was filtered. The data used to train the model for text classification was specifically about nature, hence it was necessary to filter out the observations which were conducted for built users. The number of images used for analysis was 1641; 1020 belonged to the green group and 621 to the built group.

Location points

Users were tracked while inside green spaces. In the app 949 green spaces were mapped, falling within 760 geofences. First, the location data points falling within the circular geofence but outside the actual green spaces were filtered out, while trying to avoid excessive filtering (*e.g.* people walking along the paths surrounding the green space were kept). To select the location, points with an accuracy lower than 10 meters were selected, including edge cases.

The location data was used to infer the dwelling time in the green spaces. For the time analysis, only the green spaces contained within the 5 kilometres radius circle centred in the Sheffield city centre were counted, as shown in Figure 2.2. This resulted in 539 green spaces being analysed, corresponding to approximately 78 square kilometres and 1184702 location points.

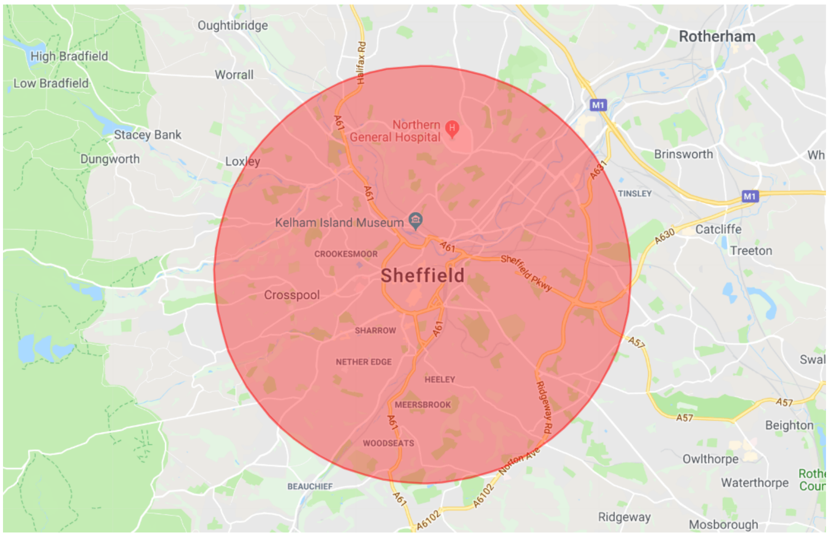


Figure 2.2: The study area for time analysis.

2.3.4 Text analysis

The comments uploaded by users are obviously personal and differ in nature and content. In the first phase, the K-means clustering algorithm was used to create groups using the dominant terms in comments. A number of K = 40 clusters was obtained experimentally, to fit the distribution of the observations themselves (created using the Euclidean distance between sentences obtained from a similarity measure). Gathered in 40 groups, the labels of the study on human connection with nature conducted by [26] were used to cluster all the observations received into 11 themes using content analysis, a systematic technique used to encode large volumes of data [40, 41]. Table 2.1 shows both the themes and a brief description, and the distribution of the theme within the dataset. The association of the observation to the theme was made through the Fasttext API which solves the [42] multi-label problem. Both labels and probability are then extracted and only those that have an accuracy of at least 50% are considered.

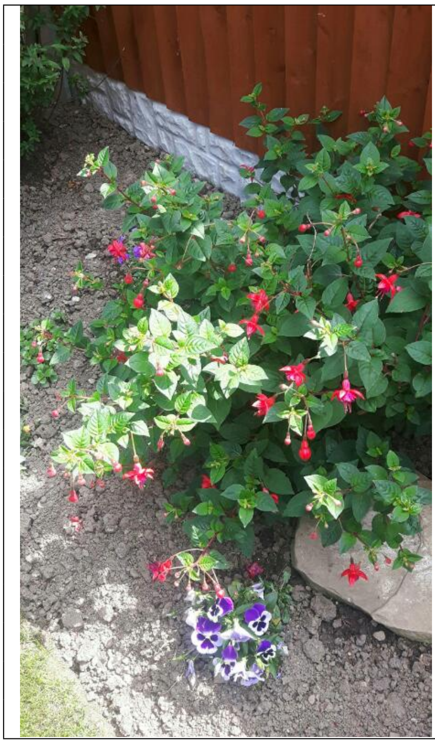
Table 2.1: Labels from training data [26].

S/N	Theme	Description	Example	No. Samples
1	Specific part of nature	When an example of a specific plant, animal or feature of nature was given with no or very little context	A bumblebee; Bluebell wood; Bright rainbow; Beach	100
2	Animals being active in their habitat	When animals were discussed in terms of some activity in their habitat	Pigeons walking in a group together like a family; A buzzard being mobbed by crows; watching 2 birds dance together; Squirrels running up a tree together	109
3	Animals interacting together	Reference to animals engaging in an activity with at least one other animal such as playing/chasing/hunting	Pigeons walking in a group together like a family; A buzzard being mobbed by crows; watching 2 birds dance together; Squirrels running up a tree together	47
4	Sensation of nature	Items which focus on the sensations of nature; smell, sound (including bird song) or touch.	Sun on my skin; Birds tweeting in the trees; Sound of long grass in the wind; hearing the birds singing to one another	159
5	Colour	Items which had a specific emphasis on colour	Bright pink blossom on the trees; The slug that I removed from my sage plant had quite a fetching orange belly	76
6	Effect of weather on something	When the weather has an effect on a plant or another aspect of the environment	The breeze in the trees; Sunlight streaming in through my window; The long grass on the bank if the stream had been flattened beneath the weight of the raindrops hanging from it this morning	93
7	Growth/temporal changes	Reference to new buds, things in coming into bloom and changes associated with the seasons	The soft new leaves emerging on our beech hedge; Purple flowers starting to bloom; Budding leaves on the trees outside my window at work; Regeneration across the seasons	124
8	Reflections on the weather	Judgement/observation on the weather or a reflection on the dynamic weather	How nice the weather was; dramatic hail storm this morning; The constantly changing weather, from rain to bright sunshine and back	72
9	Beauty/ appreciation/wonder of a particular landscape or aspect of nature	Items which refer to beauty or a specific landscape the person appreciates. Expression of the wonder of nature or the resilience and diversity of nature	The beauty of a magnolia tree in someone's garden; Mist shrouding the trees first thing in the morning; Cow parsley in the grass verge lining the road for miles on my way home	98
10	Good feelings	Reference to nature creating positive feelings or state of mind	Walking by the brook at university was very peaceful; The sun was shining, walked past the park, everyone was smiling	40
11	Other	Statements that didn't fit into themes but didn't form a theme of their own	A nice house made of wood. The beautiful wood texture and its functions are so great; The threat of rain in the air	20

2.3.5 Image analysis

For the analysis of the images uploaded by users, the Google Cloud Vision [43] API was used. Through these APIs, the content of each image has been classified by returning its labels and accuracy, an example of classification is shown in Table 2.2. Subsequently, an analysis of the labels was carried out, eliminating the redundant ones or the composite ones and evaluating the number and frequency of labels present both for the control group (built areas) and for the test group (green areas). For composite labels that contained another label within them, the shorter label was kept, considering it as the "root" label. An example can be seen in the Table 2.2 where the labels "flowering plant" and "annual plant" contain the label "plant" and which have therefore been eliminated. There are also labels with similar meanings that could be further compressed through meaning analysis using specific dictionaries (such as, for example, "WordNet") or by manual categorisation. In this work, this step is not applied.

Table 2.2: Example of labelling for an image.

	plant: 0.98 flower: 0.96 flowering plant: 0.89 flora: 0.79 garden: 0.77 shrub: 0.75 annual plant: 0.69 herb: 0.67 groundcover: 0.65 yard: 0.59

2.3.6 Time Analysis

One of the goals for the present analysis was to compute the time spent by the users inside the green spaces. In order to achieve this, further filtering of the location points was required. The procedure used is described in the following. First, all the points inside the green spaces were selected, discarding which were outside. Then for each area and for each user, a check was done about if two consecutive recorded points in a day were created within a time limit of five minutes. It has been assumed that if two consecutive locate points are farther apart in time, it would be incorrect to consider that the user spent time there. This is because while in a green space, a user's location should be continuously recorded within a small time span. Furthermore, some parks are very small and the crossing time is very low, requiring a threshold for the minimum time distance between two consecutive location points. An example for this is Dial Way Garden depicted in Figure 2.3 covering an area of 37 square meters.

The five minutes imposed check helps us to correctly discriminate user presence in the determined park. The points satisfying the imposed condition were considered relevant. Based on this, the associated counters are increased that detect the time spent by the user within that area, the number of visits within the different parks and the number of days in which users were tracked. The procedure is repeated for each user within all the areas considered. This made it possible to obtain the data for the entire time elapsed. After this phase, data was grouped and filtered to obtain different overviews, such as the total time spent by users in the different areas, the parks with the most time spent inside, etc.

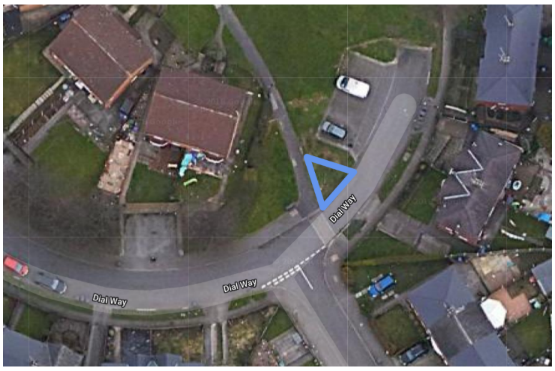


Figure 2.3: Dial Way Garden: one of the smallest green spaces in the study.

2.4 Dataset characterisation

This section presents information about the users participating in the research. In the first questionnaire provided by the app at the beginning of the trial period, the user is asked to fill in a form to have the background for each of them in terms of wellbeing and demographic characteristics. An overview of all this information is reported in the following sections.

2.4.1 Participants' description

The age distribution of the study participants is shown in Figure 2.4. The distribution is immediately unbalanced to young age, given by the massive participation of students in the study. The age range of the participants is between 18 for the youngest user and 72 for the oldest user. The rest of the work focused on three age groups consisting of young people (ages 18 - 35), middle-aged people (ages 36 - 53) and the elderly (ages 54 - 72). Each group has an equal width in terms of age (18 years). Having different amounts of participants in each group, all the results were normalised to be able to have comparable results and reduce the distortion of the data among the different groups. Table 2.3 illustrates the gender distribution of the participants. Again, in the carried analysis the results were normalised as to have a fair comparison between the two categories.

Table 2.3: Gender distribution.

Gender	**No. of users**	**Percentage**
Female	894	64.64%
Male	489	35.36%

2.4.2 Participants' interaction with Shmapped

One of the questions the users had to answer when prompted was "who they were with". Overall, 5626 entries were taken into consideration. The distribution of social interaction types is shown in Figure 2.5. It is indicating that the majority of participants were either alone or with "friends & family". The group "Other" comprises also free-text responses or a multiple selection, of which the most common was "with friends, family or partner" and "pet".

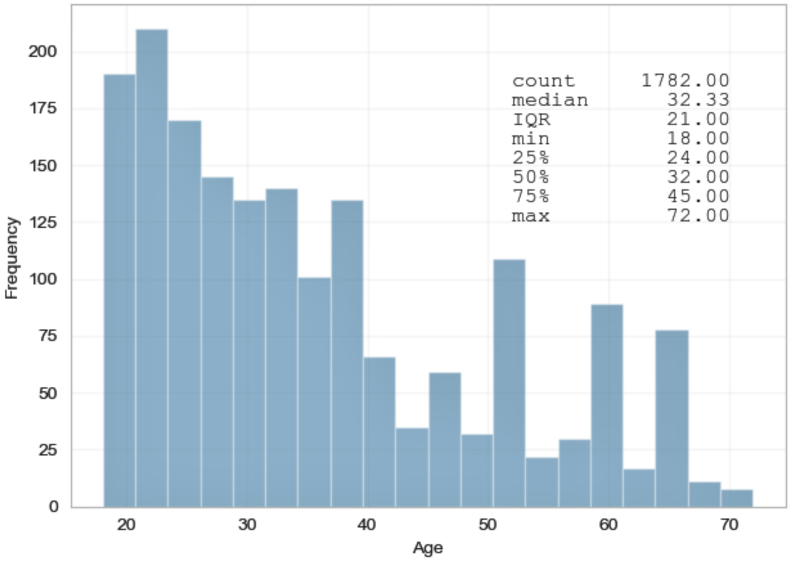


Figure 2.4: Age distribution of the sample dataset.

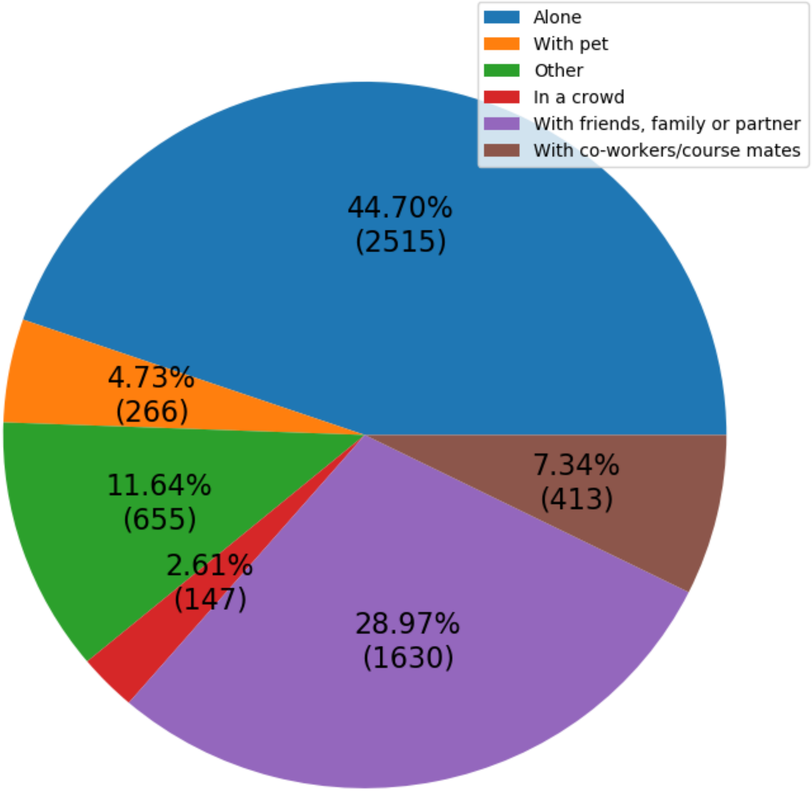


Figure 2.5: Participants' companionship / social interactions.

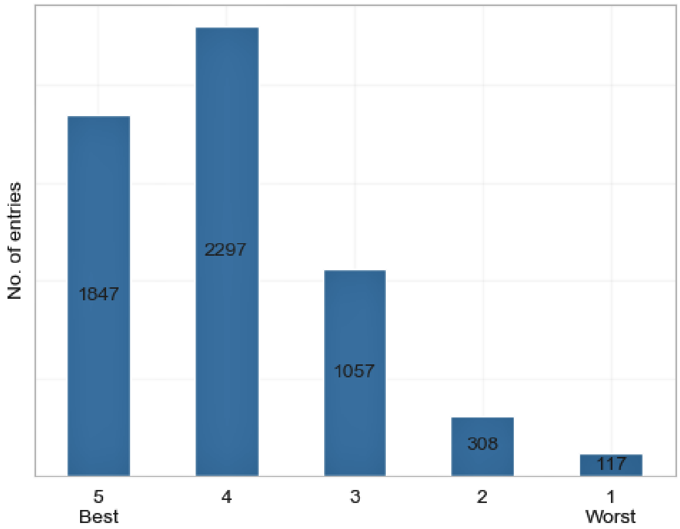


Figure 2.6: How the participants felt while interacting with their surroundings on a scale from 5 (positive) to 1 (negative).

Besides the question regarding social interaction, the users were also grading their interaction with the surrounding environment, namely how they were feeling in the situation. The histogram in Figure 2.6 shows the aggregated answers. It can be noticed that the interaction was mainly positive. Figure 2.7 shows a part of the area under examination, also including suburban parks. It shows the density of the grades they assigned. The feelings of the participants are represented by using a colour scale that varies from blue (medium) to red (high).

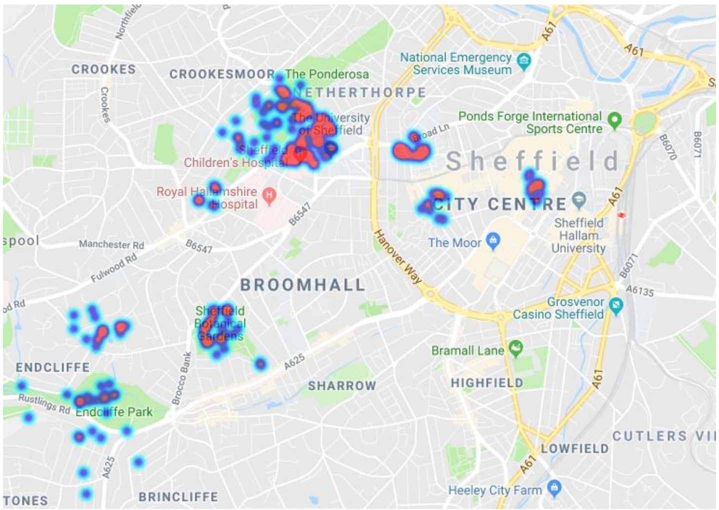


Figure 2.7: Heat-map represents the density of the users' feelings and the associated grades. The scale varies from blue (medium) to red (high).

2.4.3 Participants' wellbeing

Participants' wellbeing was evaluated based on a specialised psychometric scale which quantifies the response for each item on the scale. As a result, each user has an associated wellbeing score. The number of participants was restricted to those who completed the initial and the after-study questionnaires. As a result, the number of users decreased from the initial of 1870 to 403, because the participants either chose not to complete the after study form or they disengaged with using the app. To give an overview, for the 403 participants at the beginning of the study, 22% had wellbeing classed as clinical cases, while the rest of 78% had wellbeing above the threshold. The impact of noticing the good things in urban nature on wellbeing is reported in different works, however, statistical analysis revealed clinically significant improvements in mental health for clinical cases along with significant improvements in mental health for the whole sample, demonstrating the importance of this research.

2.5 Features noticed by the users

To find out which elements of nature attract users' attention, the observations of the participants; *i.e.* the text entries and images that were uploaded to the app were analysed.

2.5.1 What do the images say?

As described in Section 2.3.5, an initial filtering and analysis procedure was performed on the images, obtaining the labels that represent the elements captured by the participants during the study. Table 2.4 summarises the total and a unique number of labels obtained before and after filtering for each of the two groups. Taking into consideration the 10 most frequent unique labels for each group, it is possible to see what the distribution was for the two groups, trying to understand if there are differences in the elements noticed by the users of the different groups. Again, the data were normalised by dividing the number of occurrences by the total number of images in the category.

Figure 2.8 shows the 10 labels most present in both groups with an indication for the different groups. The order is given by the sum of the total observations between the two groups. For both groups, among the first 10 observations there are the labels “tree”, “plant”, “sky” and “grass”, therefore being common, adding the first 10 observations of both groups, the total amounts to 16 labels and not at 20. Being in the study required to note elements of nature, for both groups the labels “tree” and “sky” are predominant and this indicates that despite the group to which nature is salient and significant for people, as well as being the natural elements easier to notice even in the city. The difference between the two groups is however also noticed when they are analysing the total labels, for the control group, there are images with labels relating to constructions such as “buildings”, “houses”, etc. For the images of the test group, on the other hand, the labels always have a reference to nature such as “flora”, “flowers”, etc.

Table 2.4: Number of labels for participant categories.

User group	**No. of labels**	**No. of unique labels**	**No. of labels after filtering**	**No. of unique labels after filtering**
Green	9610	804	8450	676
Built	5630	640	5012	530

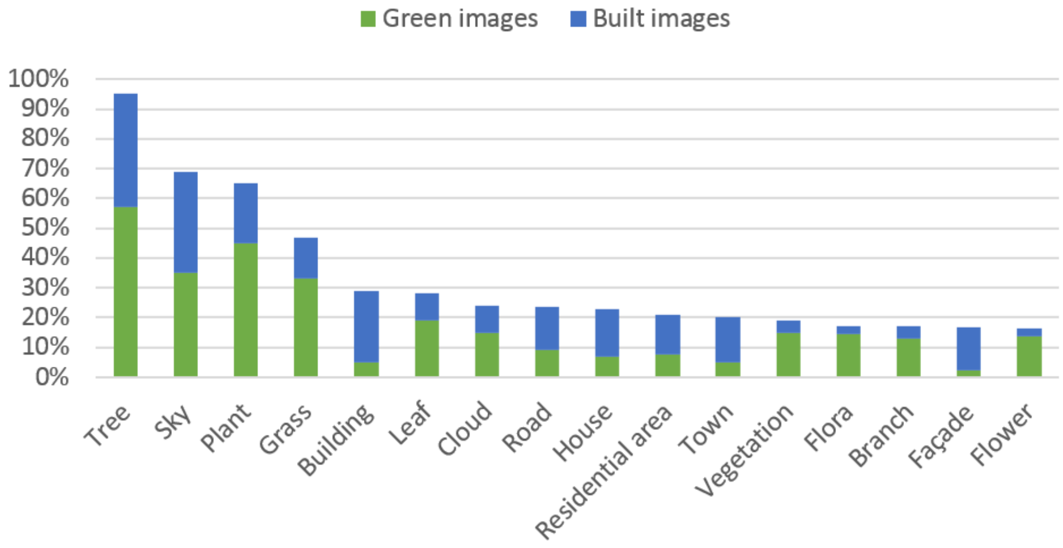


Figure 2.8: Top 10 labels for each category of images.

2.5.2 What does the text say?

As explained in Section 2.3.4 the app during the day asks users to enter observations relating to the nature that surrounds them. However, when the app requests it, it is possible to postpone the response to the evening. This specific option was meant to be a way to meet users and give them the opportunity to respond calmly at a later time. A drawback was that the position associated with the answer was stored by the app when the user entered the answer in the app and this generated not georeferenced information and for which it is not possible to understand to which green area they refer, in fact only 418 observations out of the total are recorded when they are requested by the app.

Fortunately, at least part of the observations explicitly reported the name of the area in question and it was, therefore, possible to georeference them manually. Once these corrections were made, an analysis of the texts reported in the users' comments was performed, the text clustering was performed by using a text clustering API [44]. Table 2.5 shows examples of the observations inserted, the number of observations falling in that cluster, the cluster label and the dominant term.

A subset of the clustered data is shown in Figure 2.9 where particular and different insights were obtained thanks to clustering, some observations were grouped by position, such as clusters 32 and 34 which were observations concerning parks, including some explicitly indicated as Weston Park, Meersbrook Park and Hillsborough Park. Other observations were clustered based on the type of activity they were doing, for example, clusters 0 and 7 referred to the type of walk they were doing. Interesting groups, on the other hand, concern the biodiversity of the park, including for example the terms "bird", "duck", "flower" etc. or elements of the park such as, for example, in group 4 where the central theme is the way in which the flow of the river is seen by users. Although there is a limit to inserting the observation at a later time, part of the text itself indicates a temporal component, several observations using words such as "morning" or "evening". As the study progressed, a correspondence was also found between the clusters obtained

Table 2.5: Text clustering.

Cluster	**Dominant term**	**No. Observations**	**Example Text**
1	General	1324	Shepherd wheel moss
10, 27	tree	395	Trees in nether edge
34,37	park	302	Nice park (weston park)
29	birds	195	Loads of birds in the park
20	garden	192	Insect life in our garden
0, 7	walk	186	Went for a walk to devonshire green
28	saw	158	Saw a heron in flight
2	love	152	Flowers are lovely
9	sky	126	The sky when not fully dark
26	flower	124	My honeysuckle flowers coming out
18	green	122	Green grass instead of brick or concrete greys
38	morning	113	Morning dew on the grass
13	leaves	112	Rain droplets on leaves
39	duck	93	Ducks eating carrots is pretty awesome
23	beautiful	91	Beautiful flat landscapes
19	river	68	Light dappled on the river
31	singing	55	Birds singing in the trees
15	autumn	54	I admired the autumn leaves on the trees
11	field	52	Sheep in the fields
25	weather	48	Nice weather, breezy not rainy and not too cold
35	peak district	41	Beautiful views over the peak district
12	sunset	37	The sunset when i woke up was beautiful
30	morning	36	Birds making noises in the morning
17	city	29	City centre greenery in the rain
3	heather	20	Heather covered in snow
14	snowdrops	19	Snowdrops are starting to appear
36	nest	9	Saw a nest of birds in a big tree
4	flowing	7	Fast flowing river
8	-	1	-
6	-	1	-
5	-	1	-
16	-	1	-
21	-	1	-
37	-	1	-

from this study and the recurring themes cited in the Richardson study [26].

Another interesting result is shown in Figure 2.10 which illustrates the visual output of the text classification in the eleven themes, as described in the 2.3.4 section. The "specific aspect of nature" theme was found to be the dominant theme regardless of the threshold used. For a threshold above 50%, the "theme of active animals in their habitat" is the second-highest. The top 5 themes with a probability greater than 50% in this study correspond interestingly to the first 5 themes of the study in [26] collected from a traditional and time-consuming approach to content analysis.

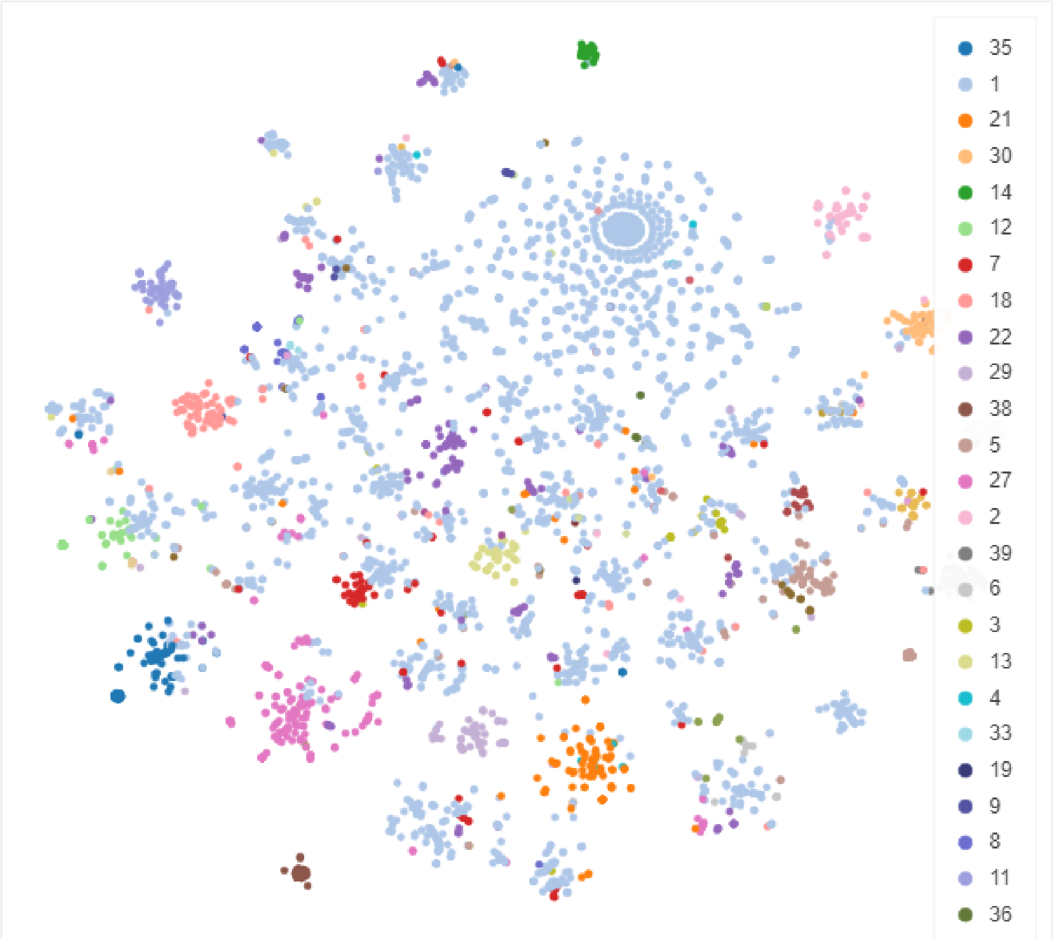


Figure 2.9: Clusters produced by k-means clustering (k=40) of textual observations. Legend captures 25 clusters.

One can state therefore that it is possible to carry out a social study of this magnitude, automating both the collection and analysis of data, optimising the study and for example being able to extend it without an additional contribution of forces. The only difference to underline is that, unlike Richardson's study which has the theme "sensations of nature" as its dominant theme, this study has as its main theme "the specific aspect of nature".

Figure 2.11 instead shows the distribution of the text classification of user comments, discriminating for each age group. Themes 1 and 9 appear to be the most popular in each group as predicted by the general ranking. For the younger group, there is less interest in the activity of animals in their habitat than in other age groups, as growth and temporal changes seem to be more interesting to them.

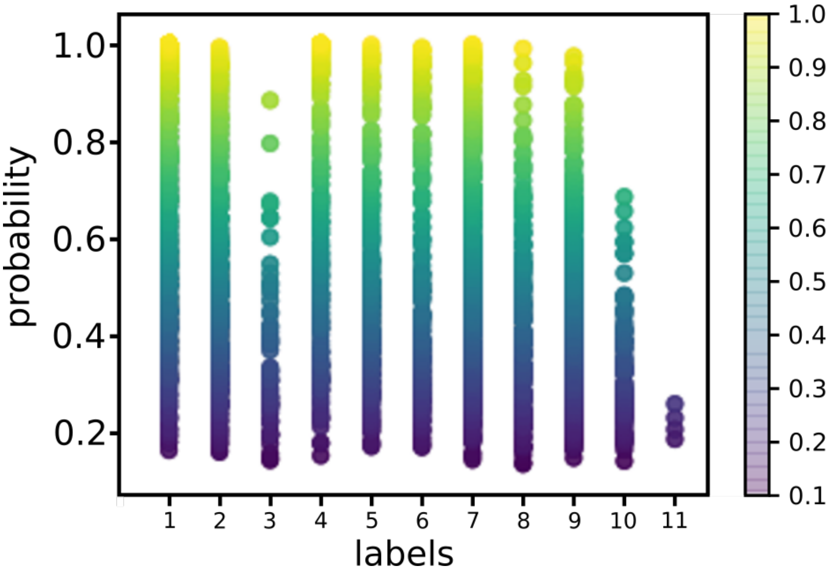


Figure 2.10: Classification of the textual observations into the themes of Table 2.1 with the FastText algorithm.

In Figure 2.12 the result of the text classification is instead reported, discriminating by gender. The results are very similar, showing a slightly greater interest in the female gender to some themes like the sensations of nature, colour and beauty. With the details obtained from this study, it is possible to understand how it is possible to use tools like these to help connect people with the surrounding environment and consequently improve their wellbeing. In fact, targeted interventions based on the sex and age of users are possible.

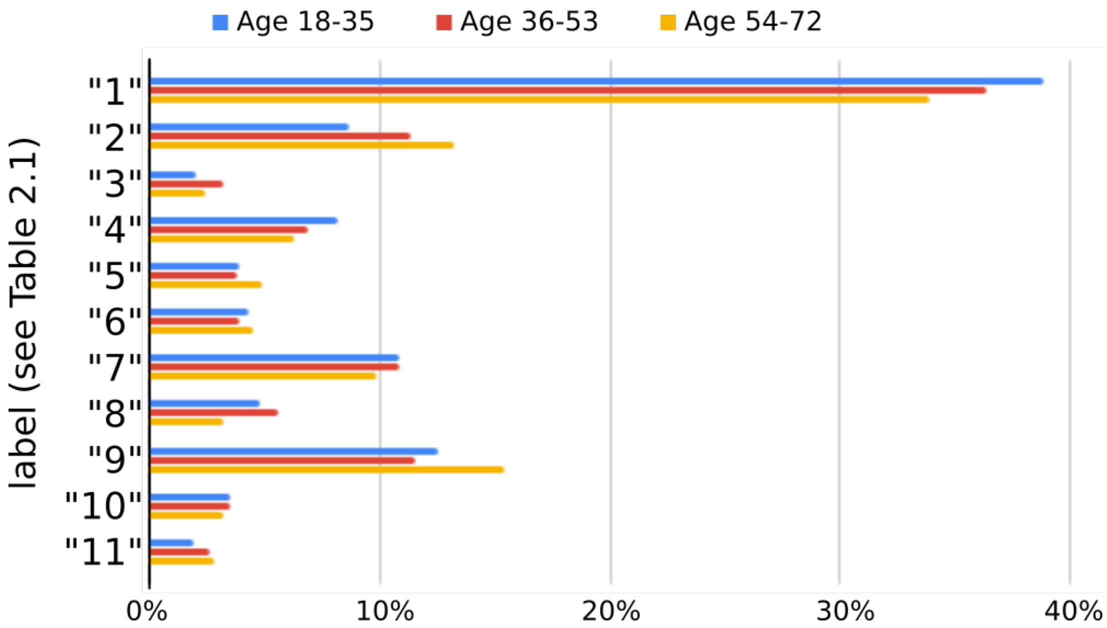


Figure 2.11: Age classification of textual observations.

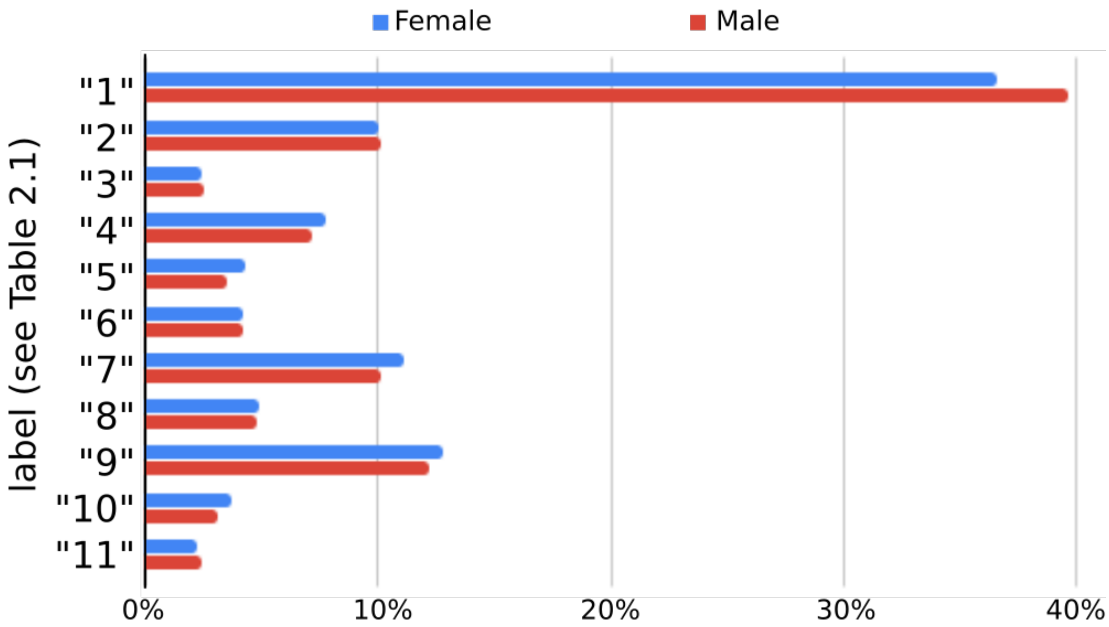


Figure 2.12: Gender classification of textual observations.

Table 2.6: Average time spent in parks by user.

User	Period Study	Tracked Days	No. Visit	Visits for Day	No.Parks	Total Time	Avg Daily Time	Avg Visit Time
1	35 Days	30	106	4	10	5 days 20:40:01	04:41:20	01:19:37
2	70 Days	23	64	3	14	2 days 00:31:09	02:06:34	00:45:29
3	38 Days	20	47	2	7	1 days 07:22:15	01:34:07	00:40:03
4	8 Days	7	25	4	6	0 days 08:17:29	01:11:04	00:19:54
5	11 Days	11	48	4	11	0 days 12:54:22	01:10:24	00:16:08
6	43 Days	19	37	2	8	0 days 21:55:20	01:09:14	00:35:33
7	69 Days	24	37	2	7	1 days 03:19:41	01:08:19	00:44:19
8	114 Days	41	168	4	8	1 days 20:00:32	01:04:24	00:15:43
9	122 Days	74	154	2	26	3 days 04:23:51	01:01:57	00:29:46
10	24 Days	16	94	6	25	0 days 16:18:26	01:01:09	00:10:25

Table 2.7: Average time spent inside green spaces by park.

Rank	Park	Tracked Days	No.Visit	No.Users	Visits for Day	Visits for Device	Total Time	Avg Daily Time	Avg Visit Time
1	Endcliffe Park	124	358	68	2.89	5.26	10 days 19:28:10	02:05:33	00:43:29
2	Whiteley Woods	71	111	23	1.56	4.83	2 days 20:24:10	00:57:48	00:36:58
3	Weston Park	149	807	170	5.42	4.75	5 days 19:31:05	00:56:11	00:10:22
4	Botanical Gardens	97	191	39	1.97	4.90	2 days 22:46:01	00:43:46	00:22:14
5	Ponderosa Park	82	231	46	2.82	5.02	2 days 00:13:41	00:35:17	00:12:32
6	Hillsborough Park	52	165	29	3.17	5.69	1 days 02:24:19	00:30:28	00:09:36
7	Hallam Square	117	287	56	2.45	5.13	1 days 09:53:09	00:17:23	00:07:05
8	Crookes Valley Park	90	246	76	2.73	3.24	0 days 23:20:11	00:15:33	00:05:42
9	St. Georges Lecture Park	109	310	76	2.84	4.08	0 days 20:05:13	00:11:03	00:03:53
10	Peace Gardens	135	334	91	2.47	3.67	1 days 00:09:48	00:10:44	00:04:20

2.6 Time spent in green spaces

In this section, the time users spent in the green spaces computed as described in Section 2.3.6 is reported.

2.6.1 Top users and parks based on average time spent in green spaces

Tables 2.6 and 2.7 offer a view of the top 10 users and parks, focused on the average time spent. It is important to note that, although the study period was defined, users were able to continue using the app for longer due to follow-up requirements. Therefore, the total time spent for different users cannot be directly compared, so an average time had to be considered in the study. For Table 2.6 the column "Period Study" presents how many days the users were part of the study, while the column "Tracked days" identifies the number of days the users were using the app and had location data recorded, meaning days in which there were associated entries.

What is interesting here is the number of parks which users interact with. Considering that the average number of parks where the participants spent their time is 7, this indicates that throughout their daily routines, people tend to interact with a variety of green spaces. Therefore, it is important to offer a high variety of parks, such as number, size and location with which citizens can interact, rather than having only large suburban parks. The average daily time spent in green space by a user is calculated as the average of all the time spent values for each day in which users have some interaction with a park.

This analysis was carried out taking into consideration only the days in which the person actually interacted with the parks, shown in the tables as tracked days, which actually means that only the days where there were location points recorded for the specific user have been taken into account. On average the users spent around 20 minutes in green spaces for every day in which they interacted with at least one park.

The top 10 users shown in Table 2.6 have a time spent in nature higher than average and, excluding user 4, also the number of parks with which users interact is higher. Then there are some borderline situations, for example, user 1 spent an average of 4 hours a day in green spaces. By analysing their data it was noticed that they spent almost all their time in a park. This suggests that they are connected to this park for a specific reason; *i.e.* it may be that they work in the park (an example could be park maintenance, a dog-sitter or fitness instructor).

The top ten parks users interact with in terms of average time spent can also be observed. Using heat maps (also called density maps) it is possible to see how the people interact with parks, where they go and what paths are the most used. Figure 2.13 shows the users' interactions with "Endcliffe Park". The heat map evolves from green (fewer location points) to red (higher number of location points). The red path identifies with the actual built path in the park which can be identified by the light coloured thin line. The green paths are rather in green spaces where there are no built paths and the users freely walk around.

It can be seen that this view allows us to identify the most used paths in a green space, as well as the less explored parts. This can act as a trigger for administration and local authorities to decide in which areas should the new interventions in that park be focused.

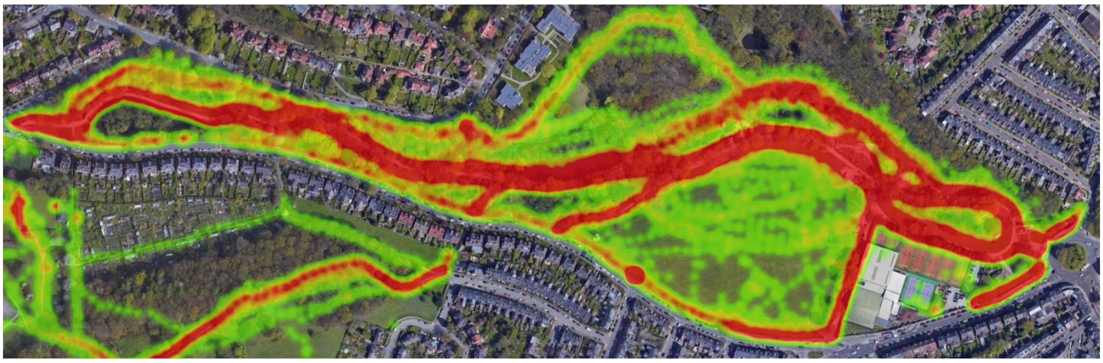


Figure 2.13: Endcliffe Park utilization based on the concentration of location points (green - low number, red - high number).

2.6.2 Age and gender distribution in park utilisation

In Figures 2.14 and 2.15 a depiction of how the different age and gender groups interact in terms of time spent with the top 10 green spaces is shown. Focusing on the age distribution, it can be noticed that the middle-aged group prefers to spend more time inside big parks like Endcliffe Park, Ponderosa Park and Hillsborough Park. Younger group instead prefers parks like Endcliffe Park, Whitley Woods, Weston Park and Botanical Gardens, while older group spent more time inside Endcliffe Park, Whitley Woods and Hillborough Park.

Analysing Figure 2.15 it is possible to assert that some parks are used in an unbalanced way by the different genders. Examples are Endcliffe Park and Hillsborough Park where there is a higher presence of male users. The opposite happens in Weston Park, Ponderosa Park and Crookes Valley Park, where there is a higher use by the female participants. This basic analysis shows how the data collection methodology could provide data of interest to local authorities and inform the design and provision of urban green spaces. More detailed analysis can explore the park characteristics and relationships to outcomes such as wellbeing.

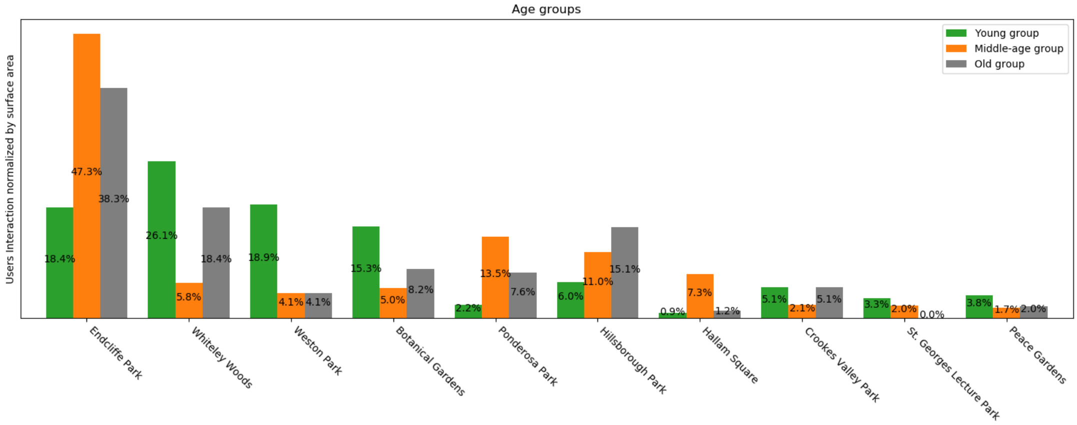


Figure 2.14: Age groups interactions with top 10 most visited green spaces. The percentages are computed only on the samples in these top areas. The amount of interaction decreased by going to the right.

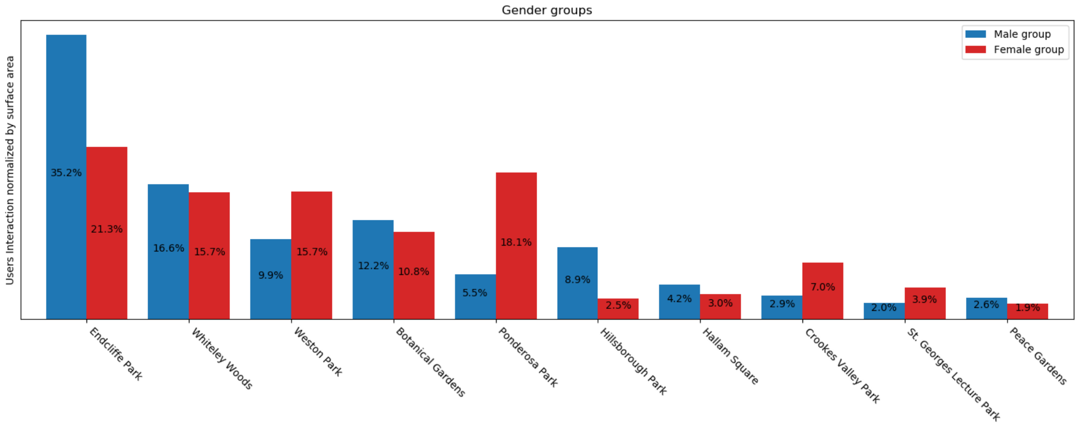


Figure 2.15: Gender groups interactions with top 10 most visited green spaces. The percentages are computed only on the samples in these top areas. The amount of interaction decreased by going to the right.

2.7 Comparison between objective and subjective interaction

In a previous work [8], the interaction of users with the green spaces through the app utilisation was analysed and a chart of the top 10 areas based on the number of observations was done. In this work a similar analysis is carried out. The interaction based on the number of observations with the interaction based on the location points is compared. The data is analysed and subdivide it according to demographic characteristics such as age and gender, so as to be able to compare the different behaviours of the users as shown in Figure 2.16 and 2.17. These graphs are based on the top parks according to overall subjective interaction (density of recorded observations).

The analysis results of location points density in these top parks are different from Figure 2.14 and 2.15 where instead the order and the data are based on the time spent inside the green areas. In these graphs the x-axis is ordered based on the total interaction density, so the interaction is higher in the first park on the left and then decreases in the parks to its right. The most interesting thing here is to notice how the subjective and objective data differ. In some cases, the users who interacted more with some parks in terms of time spent did not interact with the app in the same way.

Checking the graphs by age distribution, it is possible to state that St. George Lecture Park is actually one of the parks less frequented by the old group both for the interaction with the app and for the interaction with the park itself. In this area, the young and middle-aged groups are more consistent due to the presence of the university site. Peace Gardens appears to be a park where the interaction based on location points outnumbers the interaction based on observations in all age and gender categories. This is probably due to the area being in the heart of the city, surrounded by cafes and working spaces. Because of this, it is likely that a high number of people often pass by due to engaging in other activities such as hanging out with friends, going to work etc.

The central position, accessibility and present features seem to trigger a high objective interaction. The opposite situation appears to be recorded in South Street Park where the subjective data highly outnumbers the objective data in all age categories. Also, the features of the area seem rather different to those in Peace Gardens. Furthermore, the area is considerably larger, along a street, in an area with residential buildings. Therefore, it is more likely that people are returning home, passing by the park. This could trigger a subjective interaction as people notice green features from the distance. At the same time, the interaction concerning location points is limited as the persons do not actually go into or pass through the park.

2.8 Summary of key results

Using a real use case, related to a pilot study in the field of social sciences where the aim was to understand what was the interaction of citizens with urban green areas, it was presented in this chapter as data science techniques and machine learning can maximise the insights of the study itself. The entire case study concerned the Sheffield area, UK. The data was collected through a smartphone app that allows both subject and objective data to be acquired at the same time.

After the data was collected, it underwent several steps to return useful insights. The first phase of data cleaning (according to well established, state of the art

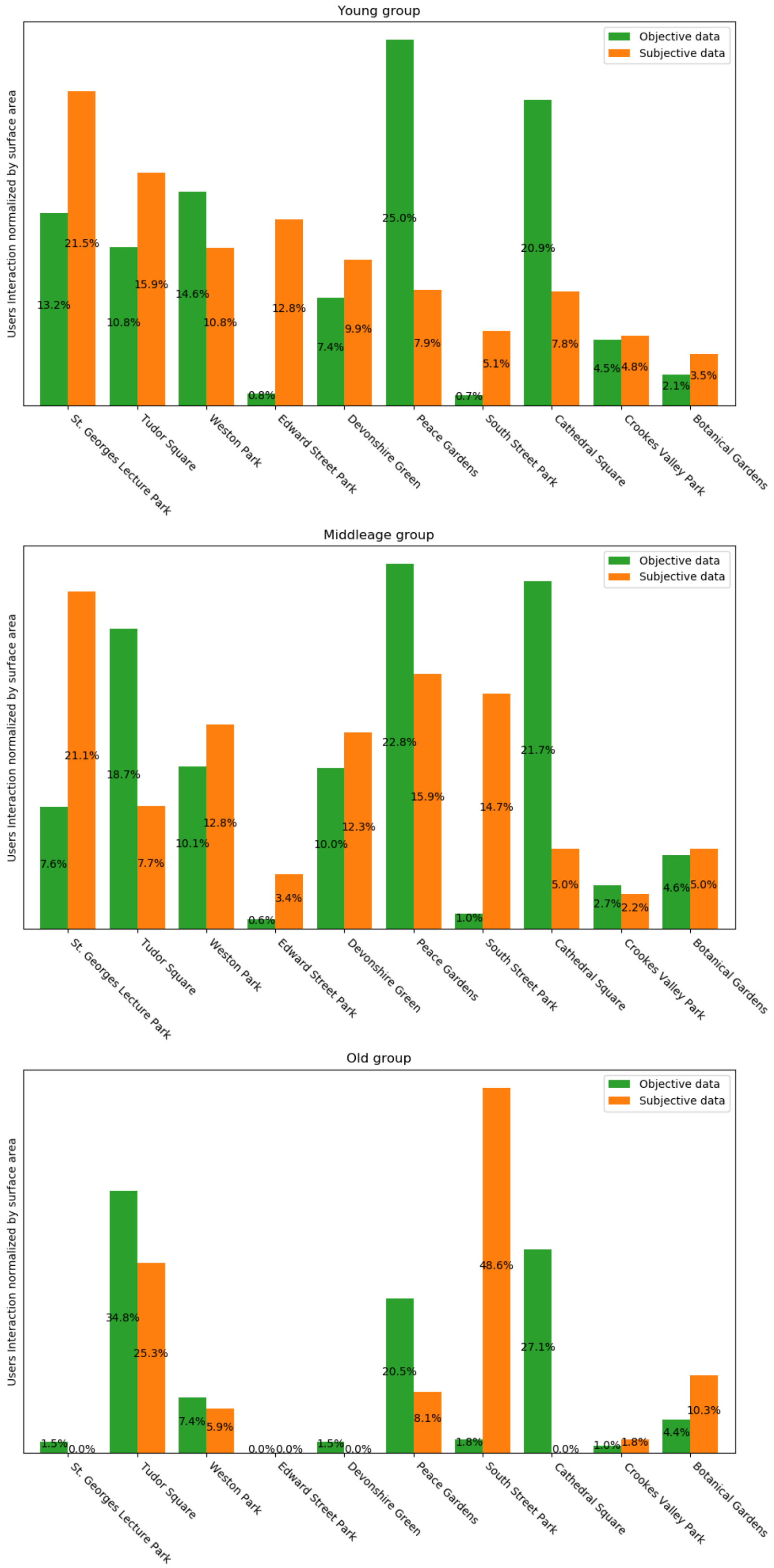


Figure 2.16: Age groups objective (green) and subjective (orange) interactions with top 10 most visited green spaces. The percentages are computed only on the samples in these top areas.

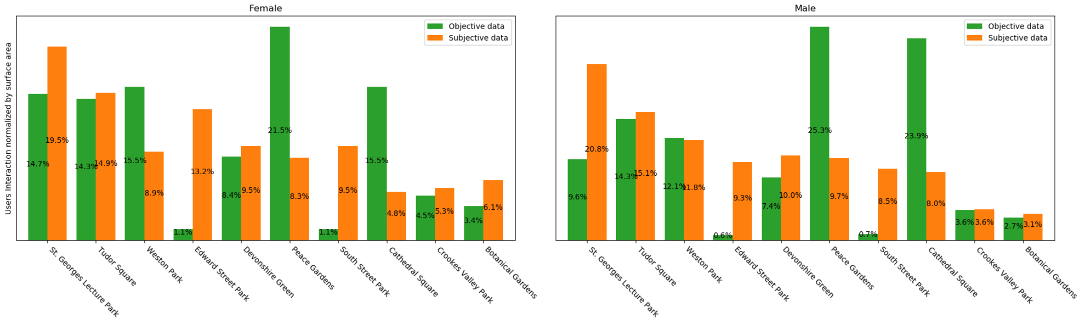


Figure 2.17: Gender groups objective (green) and subjective (orange) interactions with top 10 most visited green spaces. The percentages are computed only on the samples in these top areas.

methods) made it possible to tackle problems such as missing data, wrong data or other elements that can affect the statistical significance of the data. Secondly, once the data was cleaned and conditioned, the next phase was where the data were analysed using different metrics. Machine learning techniques were used for the texts, classifying the content and grouping the observations into 11 main topics. In this analysis, it is presented how it is possible to carry out analyses and extract insights automatically even for social studies in large areas.

The objective data was then analysed, which through the acquired localisation points made it possible to understand which are, not only the most active users but also the patterns of parks and urban green areas frequenting. Finally, the subjective and objective information processed through data fusion operations, allowed to identify connections between the variables.

Lessons learned from this work led to a better understanding of how to conduct large-scale social studies and which techniques can be employed to obtain results from both objective and subjective data. The limit of this work is that there is no type of control and addressing during data acquisition, so having questions and/or positions for which an important amount of data has been collected and others where not enough data was collected to be taken into account.

The work presented in this chapter is an improvement and help in social studies, as it allows to carry out studies also on a large scale, improving both the planning of services and the necessary targeted interventions in cities.

Chapter 3

A framework for improving subjective data collection

In this chapter, artificial intelligence (AI) is exploited to address the challenge to reduce the intrusiveness in social studies that use crowd-sourcing for subjective data collection and improve the statistical significance of the discovered insights. A method where the sensors' objective data is analysed in real-time to scope down the test matrix of the subjective questionnaires is presented. Then, subjective responses are parsed through AI models to extract further objective information. The outcome is an interactive data analysis framework for urban environments, which has been tested in the context of a citizens' wellbeing project. In the pilot study, each new entry (objective or subjective) is parsed through the AI engine to determine which action maximises the information gain: a particular question is fired at a specific moment and place, to a specific person.

With this AI data collection method one can reach statistical significance much faster, achieving a 41% acceleration factor and a 75% reduction in intrusiveness in the city-wide study. This study opens new avenues in urban science, with potential applications in urban planning, citizen's wellbeing or sociology.

These results feature in the paper "*An AI approach to collecting and analysing human interactions with urban environments*" published in IEEE Access [10].

3.1 Introduction

Progress in smart sensing, Internet of Things (IoT), crowd-sensing, crowd-sourcing, and artificial intelligence (AI) have made it possible to carry out multidisciplinary studies on a vast scale. The focus herein is on urban analytics, particularly fusing objective data collected by smartphones with subjective (citizens') responses to pursue statistical significance efficiently. To deeply understand the citizens' interaction with cities, it is necessary to attain a holistic understanding of urban environment quality [45], in relation to sustainable urban and human wellbeing development, with minimal intrusion. It is therefore essential to develop new methods to collect, analyse and fuse heterogeneous data in real-time and at scale [25], to help administrations and competent authorities in better using the public infrastructure [46], operating it with minimal budget [17].

Research in this area arouses great interest, with a pressing demand to develop new methods, algorithms and tools to manage smart cities and the complex processes around the citizens' wellbeing [47, 48, 24]. A challenging task is to manage the quality of life in urban environments through targeted interventions based on objective and subjective data, at a time when data is becoming a commodity but is increasingly hard to explore. Urban analytics inherits methods from social science and psychology to understand the human perception of the environment, measure human satisfaction, and pinpoint intervention methods [19].

It is, however, necessary to move forward from conventional methods whereby the subjective studies are based on static questionnaires, which suffer from poor statistical significance and do not make good use of the vast amount of objective data at hand. Conventional methods suffer from different problems. First, it is hard to collect sufficient samples, particularly when vast geographical areas and populations are involved, such as in megacities. What is worse, obtaining in itself the psycho-physical state of the subjects is often of little significance when taken in isolation from the context. Another common problem is the intrusiveness of tests, whereby the interviewees' commitment (and accuracy) decreases as the test time increases.

In this work, artificial intelligence (AI) is used to address these challenges, introducing a method through which the objective sensor data can be analysed in real-time to scope down the test matrix of the subjective questionnaires. In turn, subjective responses are parsed through AI models to extract further objective information. With this interactive data analysis framework, it is possible to determine (at any moment and point) which questions maximise the information gain, reducing the intrusiveness of subjective studies. In turn, the scope of urban studies could be extendable well beyond the state-of-the-art. The idea is to use an AI chatbot to mediate the interaction between the citizen and the pot of questions that want to be tested.

The chatbot is integrated in the same smartphone app that is simultaneously collecting objective data from the phone sensors. The chatbot takes into account the overall context of the study (the statistical significance of each question in tandem with the context) to decide which question to fire, when to fire it, and in which location. This represents a novel approach, as it adopts AI methods to accelerate the collection of subjective data (from the citizen), through an online fusion of subjective and objective data (*i.e.* the smartphone sensors data). This is a new way to reduce the intrusiveness of a subjective study, whilst improving its statistical significance.

The outcome is an interactive data analysis framework for urban environments that was tested in the context of a citizens' wellbeing project, which was previously carried out through static data collection methods and bulk data analysis as well as explained in chapter 2 and relative publications [8, 9, 7]. In comparison with the earlier works, herein the whole subjective data extraction process is dynamically adapting to context (*e.g.* location), objective data (*e.g.* sensor data), and subjective data (*e.g.* citizens' feedback), in such a way as to first fire those questions that lead to maximum information gain. The framework proposed in the current investigation addresses the shortcomings of static subjective studies, whose scope is generally limited by the inability to reach statistical significance at scale.

In the pilot study, the aim is to identify how urban green areas affect wellbeing, particularly the urban features that have the most impact. That requires collecting subjective responses from a sufficiently vast population, considering the extra dimensions of space and context, which would be impossible to achieve with conventional methods. With the AI data collection method, the statistical significance could be pursued much faster, achieving (in the city-wide pilot study) a 41% acceleration factor and a 75% reduction in intrusiveness. This study opens new avenues in urban science, with potential applications in urban planning, citizen's wellbeing projects, and sociology, to mention but a few cases.

The chapter is organised as follows. Section 3.2 captures the related work, for the different research topics involved. Section 3.3 introduces the proposed framework, including the key modules, and explains how they were prototyped. Section 3.5 puts the general framework in the context of a specific case study, to better illustrate the value of this approach in a practical setting. Section 3.4 gives an account of the benefits that can be achieved, providing a comparative evaluation of this proposed approach in comparison to a static data collection method. Conclusions and future work indications are finally drawn in Section 3.6.

3.2 Related work

Thanks to the significant technological advances in data sensing and analysis, it is now possible to carry out large-scale, multidisciplinary studies aimed at the interaction between humans and cyber-physical systems [49]. This can substantially aid local authorities in finding new ways to improve the utilisation of the public infrastructure and human wellbeing [17]. Ultimately, it is fundamental to explore integrated approaches to managing data-intensive systems, in planning and decision making, to offer better services and quality of life [47]. This is typically the goal of social science studies that are, however, broadly based on static questionnaires and typically suffer from poor statistical significance [19]. To counter these limitations, research has been increasingly focusing on automated data collection (*e.g.* through smartphones) and intelligent methods (*e.g.* using chatbots).

3.2.1 Chatbots

To reduce the negative effects that digital data collection platforms have on subjective tests, the interaction with users should be as simple and fast as possible, so chatbots (and particularly intelligent chatbots) have increasingly been used. In [50], Ghose et al. use a smart chatbot to answer students' frequently asked questions. This technology is also used in the medical field for different tasks. In [51] Svetlana et al. introduce a chatbot that helps elderly people in the process of recalling past memories. The chatbot in [52] connects to diagnostics tools to formulate preliminary questions to check if hospital admissions are required.

In social sciences, an app named Mappiness is used as an intervention tool to improve happiness as an element of wellbeing [53]. The participants are invited to report their wellbeing at random times of the day, while their position is constantly monitored. Urban Mind [39], is another app designed to examine in real-time how exposure to green spaces affects mental wellbeing. Starting from the two previous apps, Shmapped [54] was developed, with a double objective. On the one hand, it wanted to be an intervention tool to improve wellbeing, by encouraging people to interact more with nature. On the other hand, it was a data collection tool useful for research. It is clear that the interaction of chatbots is still not fully effective. Currently, chatbots do not have the ability to correctly handle conversations based on the social context. In [55] Augello et al. propose a chatbot model that can choose suitable dialogues, according to what the sociological literature refers to as "social practice".

There is a fundamental difference between existing literature and the proposed approach. Typically, data analysis is a post-collection mechanism, with all data being analysed in bulk. By contrast, here the data analysis is carried out online, as it is collected. Thus, while the state-of-the-art is aiming at collecting as much data as possible (for big-data processing), here the data is analysed at run-time with the aim to guide the following steps of data collection. By fusing data in real-time, the aim is to gain information at each step. This has multiple benefits: 1) reducing the circulation of redundant (unnecessary) data; 2) accelerating the process of gaining statistical significance of data, and 3) reducing the intrusiveness of technology during the subjective data collection. In large-scale social studies, this is the first time in which this type of approach is adopted.

3.2.2 Analytic Frameworks

The Internet of Things is a prominent framework for the collection of heterogeneous sensor data, which offers a unique opportunity to develop smarter and more efficient cities. City-data may feed to planning, management and decision-support systems, revolutionising the way cities operate [56].

There are currently several frameworks for analysing smart city-data. An example is "CityPulse" [57], which can perform semantic discovery, data analytics, near-real-time interpretation of large-scale IoT data, and social media data streams. Another interesting study has been done in [58], where the framework can perform real-time processing on data collected from different applications, to help in real-time decision-making.

State-of-the-art frameworks typically focus on data analysis, but there is little attention to the issue of how to collect data at scale. In [59] Datta et al. introduce an efficient IoT framework for smart cities, which offers mechanisms to mitigate a variety of smart city challenges, using cooperative crowdsensing coupled with a data-centric approach. Another interesting framework is presented in [60], which analyses the level of waste in the city waste bins (equipped with sensors). The system maintains a prediction model, which determines the optimal route for waste collection.

Existing methods tend to be strongly dependent on technology-specific solutions to improve automation and process efficiency. Yet, they use objective data, mostly from homogeneous sources, but fail to capture the citizen's point of view and quality of experience (subjective data). This is in fact the aim of this work, where the goal is to extract as much subjective information as possible, putting it in connection with objective IoT data.

3.2.3 Complexity of data collection

Most definitions of big data in urban studies are limited to the "volume" attribute. A simplistic, yet widely adopted definition of "big data" refers to any amount of data that cannot fit into an Excel spreadsheet or could not be archived in a single machine. For example, the study [28] analysed half-million waste routes to identify inefficiencies in the collection. In [29], Anantharam et al. analysed the text streams included in 8 million tweets in the San Francisco metropolitan area.

In this study, although the aiming is not to reach such levels of data volumes, there are other difficulties typical of big data sets. There is a significant variety of data, ranging from objective to objective data, and including both structured and unstructured data. Another challenge to tackle is the data variability, since the interpretation of similar data values is sensitive to the context and time in which it is collected. There is also to deal with data uncertainty and bias, particularly, since subjective data is being used [30].

3.2.4 Objective and Subjective Data

Most technological developments have been focusing on automating and accelerating the collection of objective data, such as those coming from sensors, smartphones and other IoT devices. Collecting subjective data in a reliable and statistically significant way poses serious hurdles. An example of objective data collection is presented in [31], performs users' activity detection through wearable accelerometers and, in turn, adopts gamification to encourage physical activity. Another interesting pilot study is presented in [32], an urban mobility project based on real-time traces of both traffic conditions and pedestrians. The data is objective in nature, being collected through GPS, smartphones, buses and taxis.

The importance of computational inference on objective data is highlighted in the literature. Particularly, the interaction among intelligent objects and humans is crucial in the study of Smart Cities [34]. Yet, objective data is still affected by unreliability, inconsistency and uncertainty, for instance, in connection to sensor accuracy and missing data due to faults or communication issues [33].

On the other hand, subjective data collection (as in this study) has been targeted somewhat less frequently. In this case, the problems are:

- collecting subjective responses is less prone to automation;
- it is difficult to collect statistically sufficient data;
- data is inherently unreliable and suffers from human bias.

As matter of fact, social networks have made it possible to collect subjective data, such as tweets about local events [35]. Yet, social networks data is not immune from errors, bias and uncertainty, especially when data is subject to interpretations demanded by inferential engines. In the context of smart cities, the process of collecting subjective data can make the analysis richer, more diversified and complementary [36]. And this is the specific target of this investigation.

In all the related work the data is acquired, analysed and used to extract insight and have a better understanding of somewhat. None of them cares about the amount of data gathered or the amount of questions sent through a chatbot to people, in order to achieve responses. This is exactly the shortcoming in previous work that this chapter wants to address.

3.3 Proposed Framework

3.3.1 Overview

The proposed framework realises an information gathering and analysis tool, with emphasis on real-time collection of objective and subjective data in urban environments. The proposed tool relies on a typical software architecture, based on client-server communications (Figure 3.1). No specific libraries were used, since all necessary routines were custom-made, as will be detailed in the following sections. The client devices are smartphone apps, equipped with intelligent chatbots whose key goal is to handle a seamless interaction with the citizens. The subjective questions fired by the app are first pondered by the chatbot to minimise annoyance and maximise the information gain achieved with each question.

The system is designed to be generic and adaptive to the context through simple customisation. The chatbots are provided with a list of questions (collectively, the subjective questionnaire under investigation), along with the geographical boundaries of the study, or “entities”, which may include also specific points of interest. Examples in this study are green areas, buildings, roads, parks and so on. A better understanding of human interactions with these areas and what the effects these interactions have on well-being are the main goals of this study.

All the information collected by the chatbots is collected into the server, which has sufficient resources to store and process data, fusing subjective data with objective data, and contextualising the statistical significance of each data point. Through data fusion and inference, a feedback is provided back to the chatbots, in a way that further trigger points and questions can be raised. In this way, the questions that are fired by each chatbot are those that are associated with maximal information gain. Figure 3.1 gives a snapshot of the chatbot interaction diagram. Figure 3.2 shows an example of the textual interaction between chatbot and user.

3.3.2 Client side

Although each chatbot is fed with exactly the same set of subjective questionnaires, individual questions are fired at different times, picking first the questions that have minimum statistical significance. This, in turn, will depend upon the context of the question, considering both the internal context (within the individual chatbot) and the external one (an aggregate of all contexts experienced by the chatbots across the system). Specifically, the internal context accounts for things such as the level of intrusiveness reached at a given point (how many questions the same user has been asked before); the position of the user (if they are on an area that is missing subjective responses or not); and other objective data from the sensors (*e.g.* the level of motion, type of activity, etc). On the other hand, the external context takes into account the statistical significance of each question, aggregated across all users.

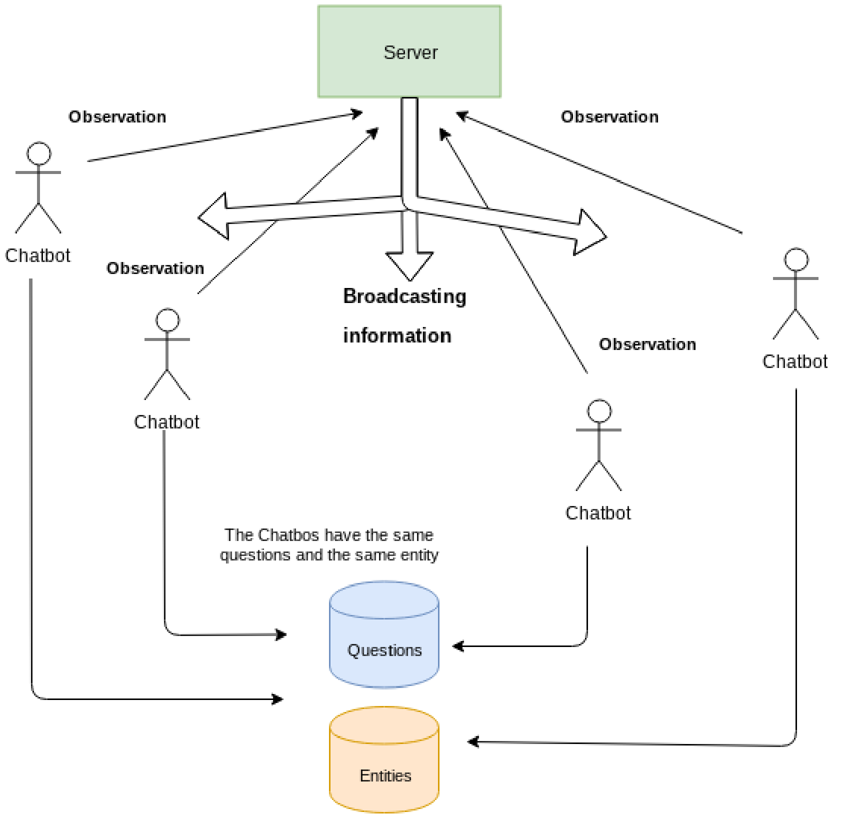
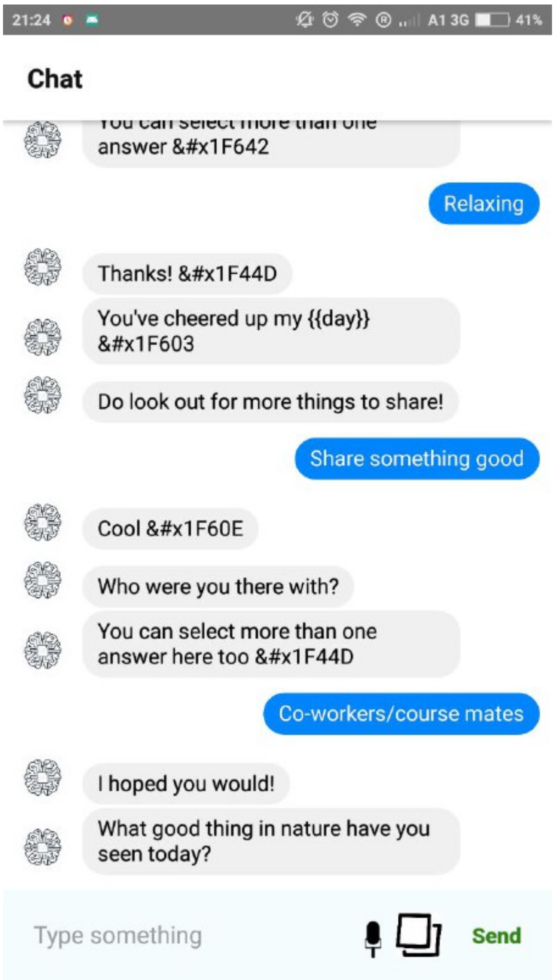


Figure 3.1: Interaction diagram between clients, chatbots, and server.

The chatbot query system emulates a normal conversation that dynamically chooses the questions to ask (the one with minimal statistical significance), taking into account the overall progress of the analysis. This methodology aims to optimise the information gain compared to what would be obtained with conventional statistical methods. Moreover, a more accurate choice of questions allows the chatbot to ask fewer questions, leading to a reduced intrusiveness of the app. At the same time, there is the added benefit of collecting subjective responses, which goes well beyond what may be achieved by making simple inference on sensor data. About issues of privacy and ethical collection of data, which always remains a critical point with subjective studies [61], it should be noted that all data is anonymized and aggregated, using it only for statistical purposes [62].

A key component of the chatbot is in charge of handling the questions. It determines which question to ask, in which moment to fire it, and which user will see it. These variables depend on both the internal context (within the chatbot) and the external one (on the server side). Three different situations may occur:



You can select more than one answer &#x1F642

Relaxing

Thanks! &#x1F44D

You've cheered up my {{day}} &#x1F603

Do look out for more things to share!

Share something good

Cool &#x1F60E

Who were you there with?

You can select more than one answer here too &#x1F44D

Co-workers/course mates

I hoped you would!

What good thing in nature have you seen today?

Figure 3.2: Chatbot Interface.

- The server communicates the questions that need to be answered;
- The server communicates a new question, generated based on the data already analysed;
- The server provides an external context, not affecting the questions order.

3.3.3 Server side

The proposed framework follows the client-server architecture and is ultimately compatible with cloud-based service provisioning. Next, the server-side overall structure and key modules are presented.

The basic server functions are:

- Storing the users' data (both subjective answers and objective sensed data);
- Analysing the data received from all clients, to extract valuable insights and make inference;
- Producing the overall (aggregated) external context, which is pushed to the individual client-side chatbots.

The whole system is organised in modules, with individual components being independent from each other and having internal functions that can be extended independently. Next, the modules that handle the webserver, data storage, data analysis, and external context management are explained.

Web server

This module is a Web Server Application, providing various services that can be reached through the https protocol via POST and GET calls. It collects users' information by means of the client-side chatbots, including both subjective (observations) and objective data. The latter includes sensory data (*e.g.* user activity) and geolocation information collected in the background. The server will also push updates of the external content to the chatbots.

Data storage

The database model used is NoSQL, which provides a flexible data model without fixed schemes that can support large volumes and the broad variety of data generated by modern applications. There are several implementations associated with the NoSQL concept. The one use is the document-oriented model, in which each record is stored as an entity with its independent properties. The framework uses two collections: the "positions" in which all the geolocation information collected by the users are stored; and a separate "user" collection in which the user's information is stored together with all other associated data.

Data analysis

As previously explained, the whole framework (developed from scratch), is built on a typical server-client architecture, in which the client-side sends the necessary data to the server through a JSON file, including only the data that is strictly necessary to make data fusion and update the contextual information. The server-side includes also a data pre-processing, integrity and verification step, in order to counter errors that may be due to transmission or data-entry errors. Classic checks are performed on outliers, missing values and inconsistent data. Data pre-processing algorithms remove those entries that have missing or incoherent data. After this pre-processing phase, the data analysis module proceeds with extracting relevant data from the acquired client-side information. This is a customisable, application-dependent feature that requires the modification or introduction of new plug-in modules. With regards to image and audio analysis, a mix of third-party APIs and custom-made components are used.

External context management

This component operates in the background, on a separate thread, to create and process the external context. The various messages received from the chatbots are first analysed through the data analysis modules. The resulting data are filtered and fused to produce the external context, as shown in the flowchart of Figure 3.3. It is important to remember that the external context is used by individual

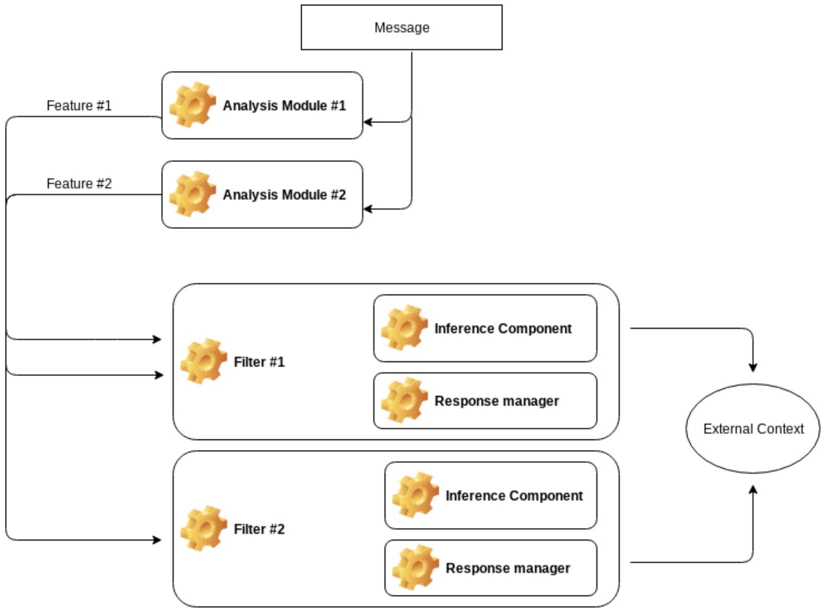


Figure 3.3: Analysis flow.

chatbots in conjunction with the internal context, to determine the questioning order, as explained in section 3.3.2.

The filters shown in Figure 3.3 are key to determining the level of statistical significance of each question (of the subjective study), and to compute the information gain attained at any moment/location by firing specific questions to specific users. In this way, a real-time status of the overall information level of the system is obtained and can turn the system towards statistically weak data. Thus, each filter contributes to the external context creation, which is saved in a JSON file that is then sent to the clients/chatbots. These can thus prioritise the questions that are statistically weaker first.

3.3.4 Data privacy and data security

While the purpose of this chapter is not directly focused on security and privacy aspects, they remain an important part to consider in framework development. For this reason, the proposed framework provides HTTPS connections ensuring that all communications are encrypted using SSL/TLS protecting the integrity and confidentiality of data exchanged between clients and server. All the acquired data is stored on MongoDB which guarantees data security by advanced security features. Some of these features are:

- Authentication: credentials are necessary to access the data;
- Authorisation: granular permissions could be defined for a user, based on the privileges it needs;
- Auditing: A native audit log tracks access and operations performed on the database;
- Encryption: all the data stored in MongoDB are encrypted.

This database's security architecture handles data security in order to defend the database against data breaches. Furthermore, no direct link can be made between the stored data and the people from whom it was acquired. In fact, even if the framework is not designed to acquire personal information such as name, surname or other, all data received by the server are pseudo-anonymised. It means that the data are processed to no longer be attributable to a specific subject. To ensure this, state-of-the-art algorithms like permutation and perturbation are applied to the data.

3.4 Internal and external contexts analysis

Recalling from Section 3.3 that the client-side receives the "external context" from the server-side, which allows the chatbot to modify its behaviour, particularly in relation to which questions to submit next, and, then, the submission rules are updated accordingly. Also, both clients and server must agree on the domain/context representation and on the users and environmental features to use. The smartphone App will have to catch all those features, as exemplified in Figure 3.4, where the domain contains the required information. The codomain elements are periodically updated by the chatbots. Some elements may be known thanks to earlier questionnaires, such as age and gender. Some others may be collected automatically through sensors, *e.g.* position and temperature. The "Question_ID" field identifies the next question to be sent to the user via the chatbot. The external context, formed as shown in Figure 3.5, is pushed to

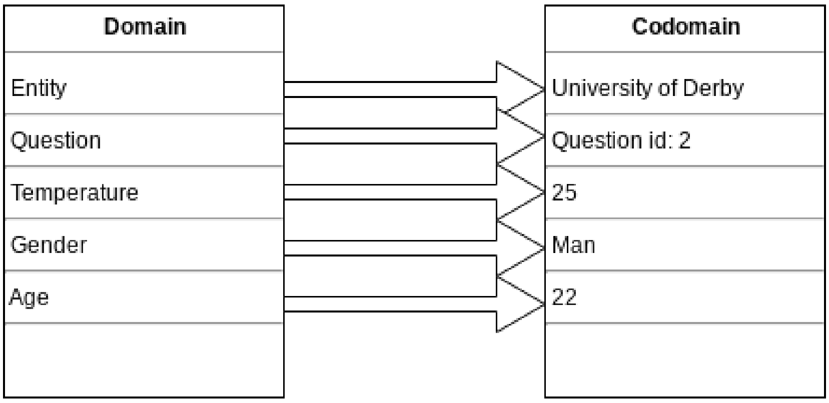


Figure 3.4: Mapping functions.

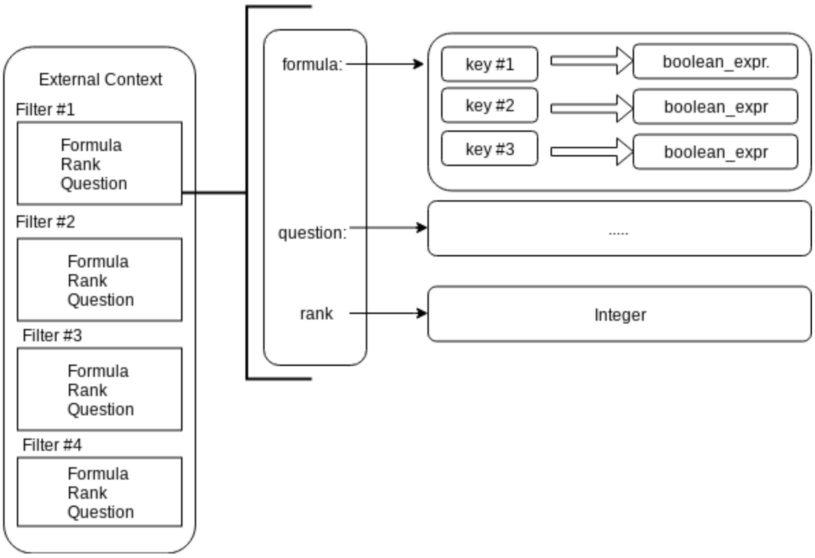


Figure 3.5: Context analysis and representation.

the client for continuous analysis. It is represented as a JSON file, containing several filters that indicate the conditions by which specific questions are fired to the user. Each filter has three fields: the "Formula" defines the conditions; the "Rank" defines the filter's priority with respect to the other filters, and the "Question" indicates the question or set of questions to be submitted to the user. This process is better specified as pseudo-code in algorithm 1. The chatbots are also managing a parameter that determines the conversation naturalness degree. Once a filter is verified, the choice of the question is influenced by two values: the distance from where the question has previously been fired, and a multiplicative factor that determines the weight of the external context with respect to the naturalness of the conversation. These two values are used together to define

Algorithm 1 External Context Evaluation.

```
for var key in formula do
    booleanExpression=set(formula[key],key,mapping[k])
    booleanResult=SafeEvaluation(booleanExpression)
    if booleanResult==false then
        return false
    else
        return true
```

the probability of submitting a question to the user. This mechanism is useful in situations where there is only one question to ask to complete the analysis because, at the same time, it is important to maintain a natural conversation between the user and the chatbot. Nevertheless, the chief aim of this study was to give priority to the information gaining process. Thus, although it is possible to play with the naturalness degree parameter to assess this aspect, in this stage this parameter was kept constant.

3.5 Evaluation through a case study

In this section, the proposed methodology and framework is evaluated, applying it to a real-world case study. The aim is to show that significant benefits, in terms of statistical significance and speed, can be achieved in urban data analysis.

3.5.1 Evaluation method

This evaluation strives for generality. An existing dataset was used, particularly one that includes a mix of objective and subjective user data, collected over a broad geographical area. The data was treated as a time series with the extra dimension of geolocation. The dataset was parsed through the emulation environment, which has the capability of replaying the very same events (*i.e.* the data collected), changing the order of the events at will. This is a key feature that allows studying the effects that a reordering of events has on information gain and, in turn, to verify the efficacy of the AI-based data collection method. It is possible to change the order and target in which subjective questions are fired, deciding the specific moment in which a question is best asked, picking specific users, specific locations, and so forth.

The emulator is a software package that relies on a database to retrieve the dataset to be replayed. Each message is taken in chronological order and analysed through the same procedures explained in Section 3.4. The mapping function is created by setting the position and the question associated with each message. Then an analytical procedure is started to determine whether the question under scrutiny satisfies both the internal and external contexts. The emulator works in tandem with the context management processes (that run in the background on the server-side system), which allows evaluating the performance of the intelligent chatbots. The key performance indicators are the statistical significance of the subjective study and the time needed to complete the study. Thus, the aim is to show that statistical significance may be achieved more rapidly by means of intelligent chatbots.

3.5.2 The IWUN Dataset

The dataset used is made up of the entries acquired during the IWUN project which was illustrated in detail in the previous chapter 2 and in related published paper [9]. However, the initial studies based on this dataset have been based on a post-collection (offline) analysis which proved not only the difficulty in terms of achieving statistical significance over broad geographical areas [8, 7], but also the benefit of urban analytics, for the purpose of understanding the interaction between citizen and city, which has motivated the new approach proposed herein (based on real-time, intelligent data collection/analysis) to pursue more effective urban analysis studies. Starting from the complete IWUN dataset, a curated version was created involving the most significant among a total of 5,626 observations.

3.5.3 Setting threshold goals

To appreciate the benefits linked to the proposed real-time (online) approach using intelligent chatbots, a comparative evaluation was carried out, benchmarking against the earlier method that was using batch (offline) analysis and static chatbots. In essence, the intelligent chatbots used the internal and external contexts to determine which questions to fire first and in which location, based on the information gain attainable. Also, the target users were chosen to minimise intrusion (*e.g.* users that had provided the least amount of feedback were asked first). In the first type of the experiments, it was set as a goal the total number of responses to collect, for each of the questions included in the subjective questionnaire pot, and for each region under scrutiny (the urban green areas). There was also set a constraint on the maximum number of responses that each user was asked to provide, to minimise intrusion.

For the sake of statistical significance, a restriction about the comparative evaluation was set to the top-20 most visited areas, chosen out of the whole dataset that originally included 760 green areas in the city. This choice was driven by the dataset at hand, which had been collected prior to the proposed (intelligent collection) method, and was found not to be statistically significant over the whole city (an insufficient number of sample answers was available, despite the scale of the pilot study). For the same reason, and for the sake of simple visualisation, another restriction about the subjective test matrix to three different questions was set, setting a target number of responses to 8 per question and per region.

Figure 3.6, shows the significant acceleration in information gain attainable with intelligent data collection/processing. The goal is reached within the first 500 interactions between the system and the user, compared to the 2,000 messages required with static chatbots. This is because of the fact that this approach can guide the question firing process based on global information, prioritising on the least asked questions/areas first, which justifies the linear information gain graph.

Figure 3.7 provides a different view of the same process, showing how the information gain evolves over time (days). Again, a significant acceleration factor is achieved thanks to the intelligent re-ordering of events. This has been computed as shown in (3.1), whereby α is the number of days that were required to reach a specific level of global information in the original pilot study. On the other

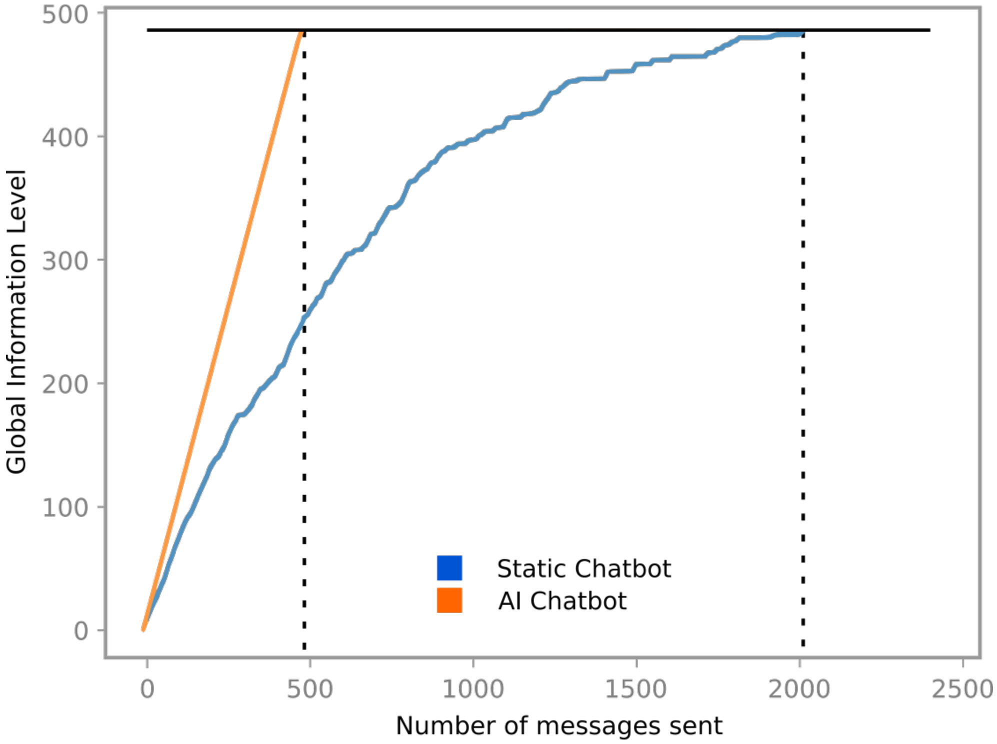


Figure 3.6: Comparative results between static and intelligent chatbots, when setting threshold goals.

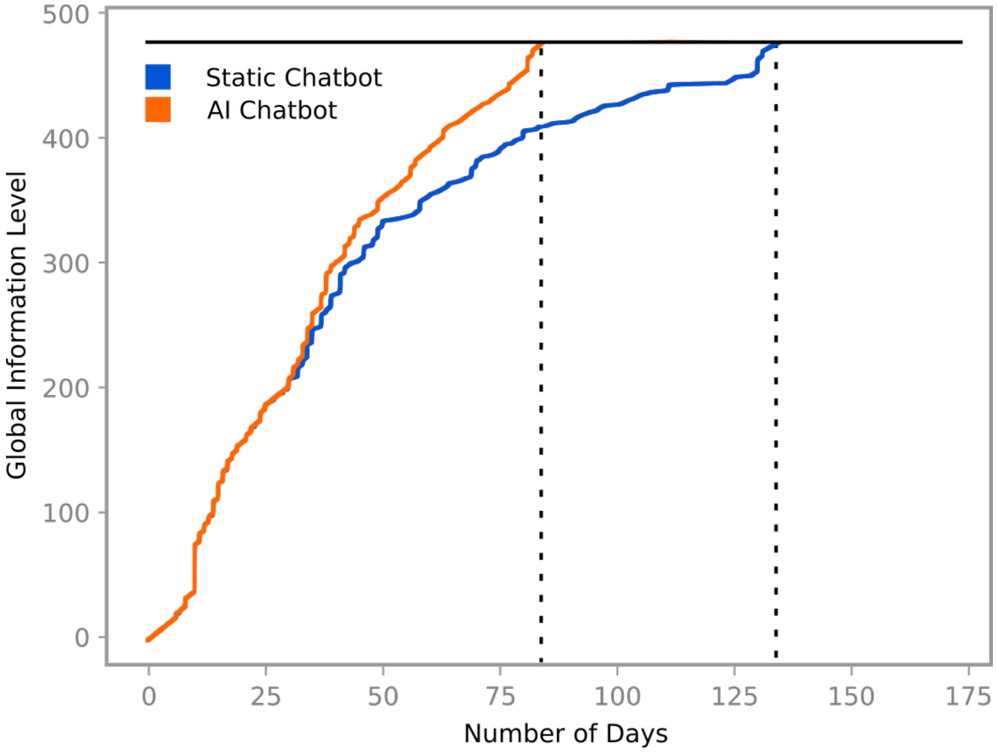


Figure 3.7: Information gain evolution over time (days), when setting threshold goals.

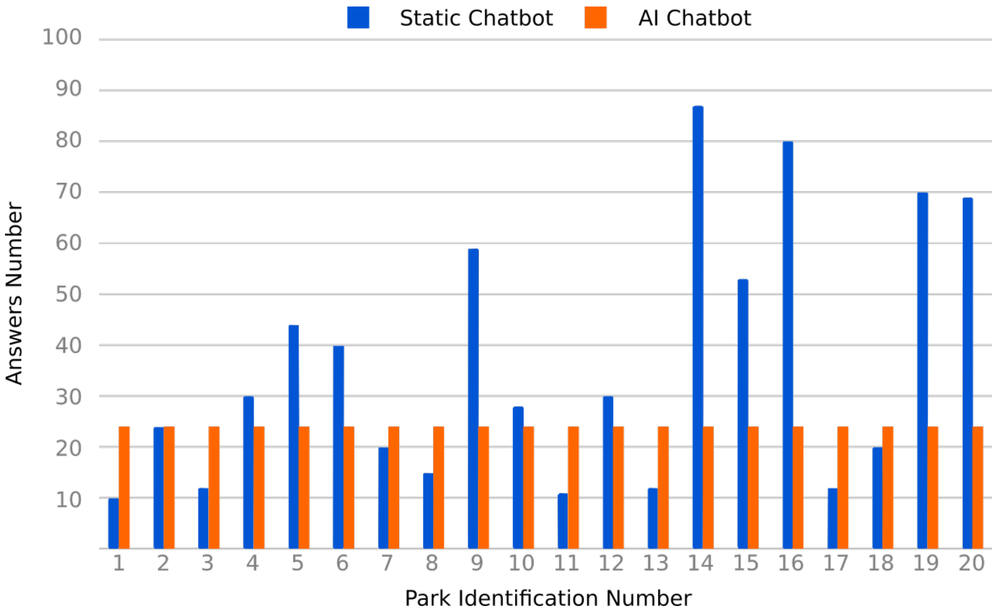


Figure 3.8: Distribution of user's feedback when setting threshold goals.

hand, β is the number of days incurred to reach the same objective through the proposed AI chatbot method.

$$acceleration\ factor = \frac{\alpha - \beta}{\alpha} \tag{3.1}$$

The spread of user's feedback achieved in the original pilot study can be compared to the re-ordered set in Figure 3.8. For simplicity, the threshold was empirically set to 25 answers (for each of the questions in the subjective questionnaire). The striking improvement (from scattered to uniform distribution) is a direct consequence of re-ordering the questions over time and location. It should be noticed that the actual threshold is meant to be study-dependent and would normally be set by the researchers of specific cases. However, the proposed method will generally lead to significant gains in information at each step (*i.e.* higher information gain per question answered).

3.5.4 Setting statistical goals

Having seen the benefits of intelligent data collection, in terms of reducing the user's interaction and duration of the experiment, the next step is to move into evaluating how far it is possible to accelerate the overall statistical process.

The goal in this context was set to reach statistical significance for each question (of the subjective test matrix), for each of the regions under scrutiny, respectively. A restriction to the top-10 most visited regions was implemented, using a set of 5 target questions.

The results included in Figure 3.9 show two step-wise functions, relating to the two methods under comparison. In this case, the information gain steps up as soon as any of the questions have received a statistically significant number of users' replies. To determine statistical significance the confidence level was observed. This is closely related to the "P value", which indicates when a specific response can be considered statistically significant with respect to the others. The confidence level parameter was set at the default value of 95%, corresponding to a value of 5% for the p-value parameter. Other typical values of 3% and 1% would influence the time required to reach a statistically satisfactory outcome of the questionnaire. Setting a lower threshold will also have a negative impact on intrusiveness since more user's feedback would be required.

This intelligent data-gathering method leads to a significant acceleration factor since both user intrusiveness and overall execution time are reduced. The intrusiveness reduction has been computed based on the formula (3.2), whereby γ is the number of messages required to achieve statistical significance in the original pilot, whereas θ is the number of messages required by the AI chatbot solution.

$$intrusiveness\, reduction = \frac{\gamma - \theta}{\gamma} \tag{3.2}$$

3.6 Summary of key results

Making insights into urban data is a daunting task, both in terms of collecting data and analysing it. The research idea was to aim at the even more complex goal of carrying out a mix of subjective/objective studies. Typically, subjective studies aim to collect relatively few samples directly from people. Conventional questionnaire-based studies are improved with digital systems, *e.g.* using smart-

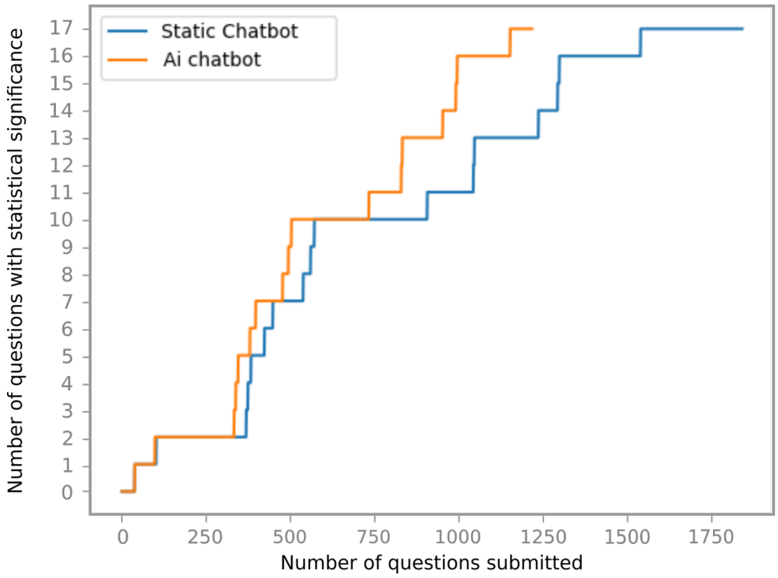


Figure 3.9: Comparative results between static and intelligent chatbots, when setting statistical goals.

phone Apps. Nevertheless, if the subjective information to be collected concerns the citizen, there is the additional dimension of space. In urban subjective studies, the users' feedback needs to be contextualised in time and space, since in many cases an answer depends critically on a specific moment and location.

This is a general problem in urban science, where optimising city processes requires the collection and analysis of subjective data (*e.g.* human behaviour, human perceptions, and human quality of experience), in conjunction with objective data (*e.g.* smart city data, Internet of Things data, and citizens' sensory data). Urbane science needs to go well beyond the analysis of objective data since most value lays with human data (and their correlation with city data). Yet, collecting subjective data at scale poses significant challenges in terms of automation and statistical significance, which is a core element in this work.

This challenge was met by adopting a use case to illustrate the scale of the problem. While the case is specific in trying to capture the interactions between citizens and green urban areas, the proposed methodology is generic. Using the IWUN project dataset, it is shown how difficult it would be to collect a statistically significant data sample in a vast geographical area. Despite being a large dataset, the IWUN data provides insufficient information to draw a complete picture, even for the relatively small city of Sheffield (UK), where 760 urban green areas have been scrutinised.

It is possible to argue that striving for statistical significance in urban science requires moving away from conventional methods, which typically separate the data collection phase from the data analysis one. By contrast, the data collection and analysis were performed, using intelligent processes in real-time (during data collection) to guide the subsequent steps of data collection. The analysis of users' feedback in real-time (through AI-based feature extraction and text analysis) and the combination of feedback with context (location and information level of each question of the subjective test matrix), lead to a significant acceleration to the overall process. In the case under scrutiny, was possible to achieve a 41% acceleration in reaching statistical significance and a 75% reduction in intrusiveness. However comparable improvements are expected to translate to other analogous cases involving both citizens and cities.

This work sets the scene for integrating intelligent data collection and analysis processes in urban analytics, which is particularly useful in urban subjective studies. Establishing when a pilot study has reached statistical significance is essential to draw reliable conclusions. A demonstration of the proposed method in the context of a citizen-to-city interaction project was presented.

Chapter 4

A passive data collection framework exploiting Wi-Fi protocol

In the previous chapters, an active data collection approach was presented and although several methods were shown to improve its efficiency, several problems remain. Among these, the number of people involved in a study remains limited and there is also the need to reward them for participating in the study. To obtain more complete objective data, increasing as much as possible the sample of people from whom to extrapolate data and totally eliminate their involvement in the acquisition phase, it is necessary to switch to a passive data collection system. In this chapter, the *People Mobility Analytics* (PmA) solution is presented, which collects probe requests generated by Wi-Fi devices when scanning the radio channels to detect Access Points. The PmA system processes the collected data to extract key insights on people mobility, such as crowd density per area of interest, people flow, time of permanence, time of return, heat maps, origin-destination matrices and estimation of people positions. The major novelty with respect to the state of the art is related to new powerful indicators that are needed for some key city services, such as security management and people transport services, and the experimental activities carried out in real scenarios.

Contributions The material presented in this chapter is a joint work between the University of Derby (UK), the Edinburgh Napier University (UK) and the University of Cagliari (Italy). The results in this chapter are based on a research initiated by the University of Cagliari [63]. This work features in the Journal of Cleaner Production as "*Pma: A real-world system for people mobility monitoring and analysis based on Wi-Fi probes*" [11].

4.1 Introduction

In the last decades, more and more people have been moving from rural to metropolitan areas. As a result, UN estimates that 55% of the world population already lives in cities and the projection shows that the urbanisation index is expected to increase to 68% by 2050 [3]. The increasing number of people living in big urban conglomerates introduces increasing complexity in the deployment and management of services infrastructure and in the allocation of the appropriate resources to reach the required sustainable urban living conditions.

Luckily, the rapid developments in Information and Communications Technologies (ICT), including Big Data, Artificial Intelligence (AI), and Internet of Things (IoT), is contributing to the Fourth Industrial Revolution [64], laying the foundations to turn cities into Smart Cities [65]. In fact, the mentioned technologies are enabling significant improvements in terms of security, people mobility, health and overall citizens life quality. One of the main applications for this kind of technology is the collection, analysis and interpretation of urban mobility data. Studying human mobility allows for making more efficient, larger-scale services, such as the public transportation service [66], the communication infrastructure [67] but also planning appropriate urban and green areas [7].

In order to get a good representation of citizens' mobility, it is mandatory to gather a large number of points, capturing people's position over time. The easiest way to collect large quantities of data with the minimum effort in terms of time and costs is to use crowd-sensing and crowd-sourcing approaches.

The main concept of these approaches is to exploit people's personal devices to extract different types of data, which can be achieved through a dedicated app installed in the user smartphone [68]. However, the main disadvantage of both participatory sensing and opportunistic sensing is that users have to play an active role in data collection because they have at least to install the app and, in some cases, they even have to provide the required input. Additionally, this approach often requires awards to be given to the involved users.

In order to properly address these issues, a viable approach is to collect people mobility data using a passive approach, which does not require users to respond actively. In a passive data collection scenario, sniffing the packets sent by devices using the Wi-Fi technologies plays an important role thanks to its low implementation costs; this is the reason why significant research effort on people localisation using this approach has been carried out in the last 15 years [69, 70].

Major studies that have been carried out so far focused on the following aspects: real-time devices localisation; trajectory tracking and people density; raw data analysis to remove useless data. Still, this research field needs significant efforts to attain practical, robust and accurate solutions. In particular, there is a need to devise the appropriate processing that, starting from the raw data, can generate the information necessary for addressing the city challenges. Additionally, there is a need to perform extensive real-life deployment to learn from the wild. The data collection module should also be respectful of the monitored persons' privacy.

To advance further in this area, the PmA (People Mobility Analytics) system was designed, developed and tested, which was initially outlined in [63], and explored in depth herein. The main focus of the PmA system is to localise people and deduce key insights about the mobility of the crowd.

Specifically, it relies on the analysis of the *probe request* packets that are sent by the user Wi-Fi devices when looking for Access Points (APs) to connect to. In this way, the devices are performing an active scanning procedure. These probe request frames contain key information about the APs visited in the past Preferred Network List - PNL - although more and more rarely [71]) and the end

device itself (*e.g.* the MAC address of the Wi-Fi interface). This study aims to determine how this information can be used to reconstruct traces of mobility, estimate crowds' density and people flows key indicators.
The primary contributions of this work are as follows:

- design and development of an architecture for a Wi-Fi based plug and play solution for people mobility monitoring and analysis;
- definitions of real-time and post-processing metrics useful to understand people habits and behaviour;
- performing extensive experiments in several real-world scenarios; *i.e.* university campus, international fairs, and roads to identify and validate suitable metrics.

The rest of the chapter is organised as follows: in section 4.2 is briefly analysed past works on people mobility monitoring and Wi-Fi localisation techniques; in section 4.3 PmA system and its components are described; in section 4.4 the procedures designed and implemented to analyse the collected data are presented; in section 4.5 the experiments that have been done with relevant results are introduced; finally, in section 4.6 final considerations about results are drawn.

4.2 Related works

The use of Wi-Fi probe requests for location analytics and people tracking has been gaining attention in the literature [72, 73, 74, 75, 76]. In this section, a brief summary of recent works, which are categorised in Table 4.1 is provided.

Table 4.1: Recent literature for Wi-Fi probe analytics.

Category	**References**
Localisation Techniques	[77, 78, 79, 80, 81]
Trajectory Tracking	[82, 83, 84, 85]
Crowd Density and Flow	[86, 87, 88, 89]

4.2.1 Localisation Techniques

Over the last ten years, a number of well-known techniques have been adopted for Wi-Fi localisation; *i.e.* RSSI-based ranging (Received Signal Strength Indicator), Time of Arrival, Time Difference of Arrival, Angle of Arrival, and so forth. However, other techniques which are in general more accurate, are also used in order to localise people, especially in indoor environments. For instance, RSSI fingerprinting, as shown by Martin et. al. [77] is a very common technique for indoor positioning and localisation. Nevertheless, this type of positioning is inefficient in urban area scenarios such as the scenario considered in this work.

A completely different approach is to extract parameters from target signals that are depending on the position of the target itself. For the Wi-Fi protocol, some of those parameters are present in the probe request frame. Regardless of the technical complexity in the implementation of a pedestrian-monitoring application, Xu et al. [78] have provided an excellent example of this type of system. Their solution uses MAC address and RSSI information, acquired by Wi-Fi sniffers, in order to localise people and to study their mobility, with the purpose of improving busses scheduling. In the same work, proper consideration has been given to the effect of environment-dependent factors like slow and fast fading.

Schauer et al. [79] figured out how to estimate the position using hybrid techniques based on RSSI and Time of Arrival information of both Wi-Fi and Bluetooth interfaces. Instead, for what concerns the arrival time and arrival time difference, several studies have been done [90, 91, 92, 93], but only recently these have been applied to Wi-Fi packets [80, 81].

4.2.2 Trajectory Tracking

As regards to the tracking of trajectories, Chilipirea et al. [82] focused on how to recognize the points where people are stationary along to a predefined path. In their work, they deployed 40 Wi-Fi sniffers during the TT Festival[1] and collected data in three editions of the festival; *i.e.* from 2015 to 2017.

[1] https://www.ttfestival.nl/

In the recent years progress with Machine Learning and Artificial Intelligence have brought enormous advantages in Wi-Fi probes analytics, as shown in [83] where Traunmueller et al. achieved good results in human mobility and human trajectories using Wi-Fi probe requests. They have used a large data-set built on the probe frames collected from 54 public APs installed in Lower Manhattan in New York, NY for a whole week. In [84], Andión et al. performed a very interesting work about Wi-Fi tracking using a low-cost infrastructure. They have monitored a University Campus that received about 4,000 people per day, during a whole year. The outcome of their work is a set of considerations about the limitations of this system, *e.g.* it is crucial to design very well the position of the sniffing stations. But the main contribution was provided by clustering methods to find characteristic behaviour of people around the Campus.

Finally, Potorti et al. in [85] presented another way to take advantage of Wi-Fi. The authors obtained noteworthy outcomes in indoor environments, such as museums and shopping malls. Without performing a survey of the environment but simply by means of existing Wi-Fi network traffic analysis and by computing the position using a trilateration approach, they have created some user trajectory into a museum and a shopping mall with accurate results.

4.2.3 Crowd Density and Flow

Wi-Fi data can be used also for crowd behaviour monitoring, as shown in [86], where Petre et al. figured out how to clean Wi-Fi data before the analysis in order to extract relevant information about the crowd. In particular, they have extracted data during an event involving 100,000 people, spread over three days.

In [87] Galluzzi et al. proposed a different approach, suggesting to analyse also the Bluetooth packets in order to improve the accuracy of people counting estimation, achieving better results in the crowd mobility estimation.

In [88] it is shown that the use of probe request information can be utilised to count people in crowds. Their contribution is provided by a device-free Crowd Counting approach based on Channel State Information (CSI). They discuss the relationship between the number of moving people and the variation of wireless channel state.

Kurkcu et al. [89] figured out how to estimate pedestrian densities, waiting times, and flows using both Wi-Fi and Bluetooth sensors. Their algorithm is used to aggregate and clean data but also to fuse additional information in order to improve the accuracy of waiting time estimation. The method was applied to a dataset collected in a transit terminal situated in New York over two months.

The common limitation of the related works presented is that they have concentrated on one or more aspects of acquisition and processing, but none of them has developed a complete framework. In some cases, the data was acquired directly from already installed access points; although this is positive in terms of the scalability of the solution, it does not solve the problem of data acquisition in scenarios where there is no pre-installed infrastructure. There are also works where the sensors for the acquisition have been described but the authors don't give any detail or attention to the architecture of the entire system, in [89] for example, the authors mention that the acquired data are saved in SQL lite but data management is not detailed, for example, which makes these solutions difficult to reproduce. In this chapter, the aim is thus to present a complete solution, from the acquisition to the metrics and insight, taking care also to the data storage, privacy, security and communication between clients and servers. In addition, some indicators like the O/D matrix and the heatmaps showing the crowd density were not introduced before in the presented scenarios.

4.3 The PmA system

As previously introduced in section 4.1, PmA (People mobility Analytics) is a complex system for collecting, analysing and interpreting data about mobility on urban scenarios. Figure 4.1 shows the two main elements of the system: the PmA Stations and the PmA Platform.

The PmA Platform deals with the collection, processing, storing and visualisation of data; whereas the PmA Stations have the task to collect and pre-process the

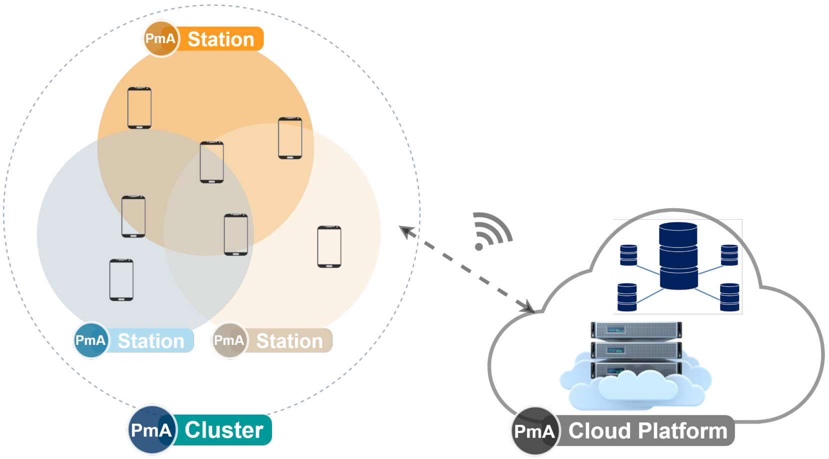


Figure 4.1: Macro-architecture of PmA system.

locally collected data. Multiple stations are grouped in PmA Clusters, so as to facilitate devices management. The PmA Platform is composed of Server and Storage components, which are located in the cloud.

Figure 4.2 provides further details about the architecture of the whole system. Data persistence is guaranteed by a distributed storage system configured in a replication set to provide high reliability and robustness. All the parts concerning the manipulation and visualisation of the data are running on the server-side, which also offers the entry-point for the system's components communication. In the following subsections, details and specifications about the PmA platform and stations are provided. Next, the privacy issues that have been considered at design time are discussed.

4.3.1 PmA Stations

PmA stations have been designed and implemented with low-cost components: Figure 4.3 shows an example of a real sensor station. Stations are composed of a Raspberry Pi3 model B+ and high gain antennas (to extend Wi-Fi coverage) and are configured in monitoring mode, which allows for packet sniffing. The specifications of the antennas used in the stations are shown in Table 4.2. Next, the different modules of the system architecture depicted in Figure 4.2 are illustrated.

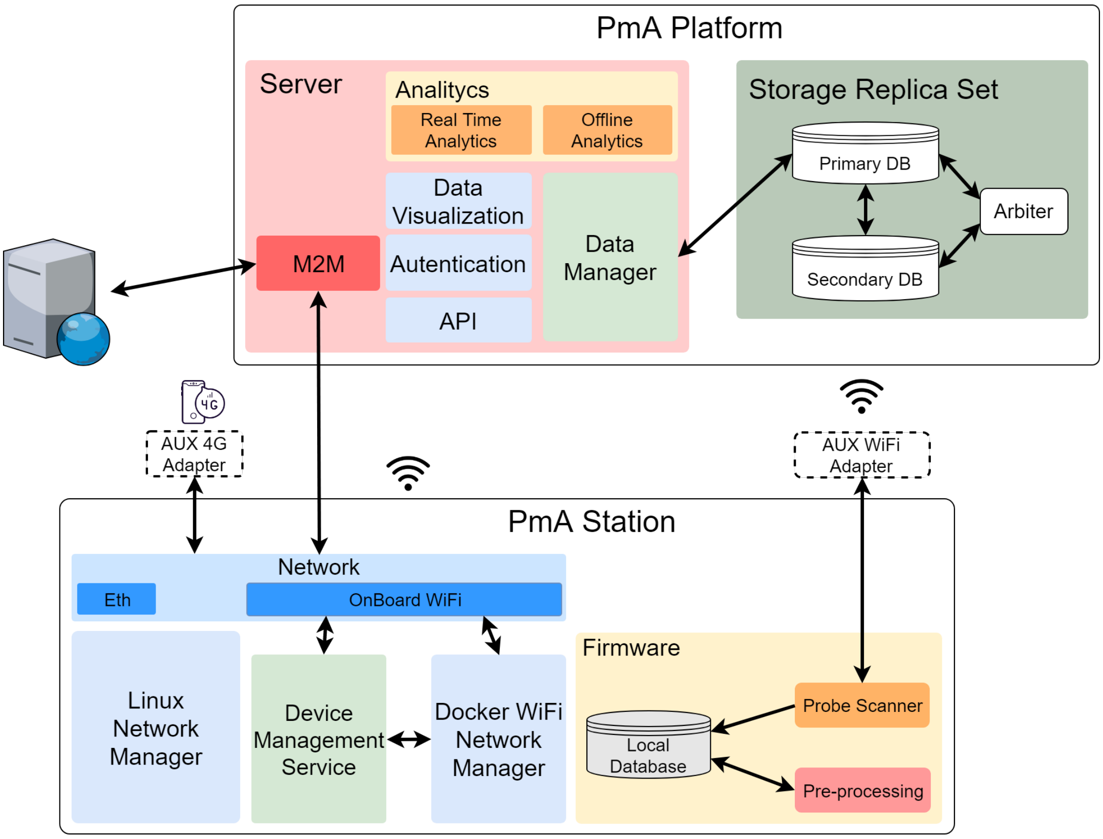


Figure 4.2: Detailed system architecture.

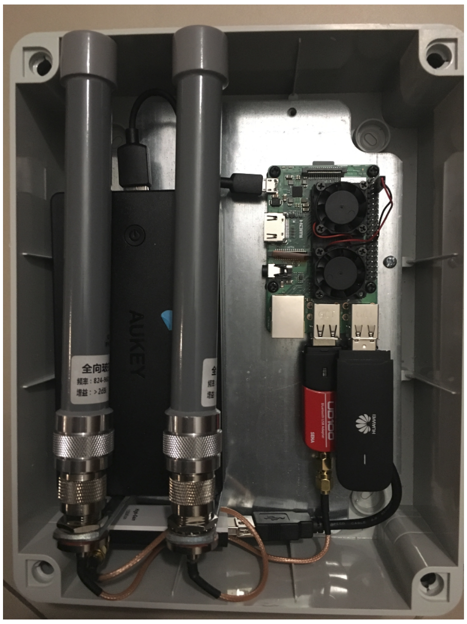


Figure 4.3: Outdoor PmA station used in the pilot study. The picture reveals the high gain antennas used.

Table 4.2: High-gain antenna details.

Frequency	2.4 GHz - 2.5 GHz
Gain	5 dBi
Impedance	50 Ω
S.W.R.	¡= 2.0

The “Linux Network Manager” module takes care of managing the Ethernet interface. Stations can connect to the network using different kind of technologies, including mobile technologies accessible via the “AUX 4G Adapter”.

The “Docker Wi-Fi Network Manager” block contains a dockerized script that has been implemented to manage the built-in Wi-Fi interface, which allows Access Point mode and connectivity at the same time.

The “Device Management Service” module is used to configure the device. The configuration interface is available by connecting to a local server which is running inside the device, such as a home router. To allow this type of configuration, a new network manager has been implemented alongside the linux OS, which is the default network manager. Thus, it is possible to set the Raspberry Pi as an Access Point.

The “Firmware” module is independent from the other modules. It scans probe requests using the high-gain antenna represented in the “AUX Wi-Fi Adapter” module. The pre-processing sub-module extracts information and passes it to the local database, to temporarily store it before it is sent to the PmA platform. For the purposes of experimentation, the PmA stations are equipped with batteries, with connectivity provided through the Wi-Fi interface.

In order to configure the station correctly, the first step is to use its management interface by means of connecting to the network created by the stations themselves. Through simple steps, it is possible to select the station’s digital twin (when this already exists), through a query to the platform. Otherwise, it is necessary to create the digital twin and then repeat the process. In both cases, a configuration file will be created in the PmA station file system. These steps take place after performing user’s authentication, to verify that the user has a valid

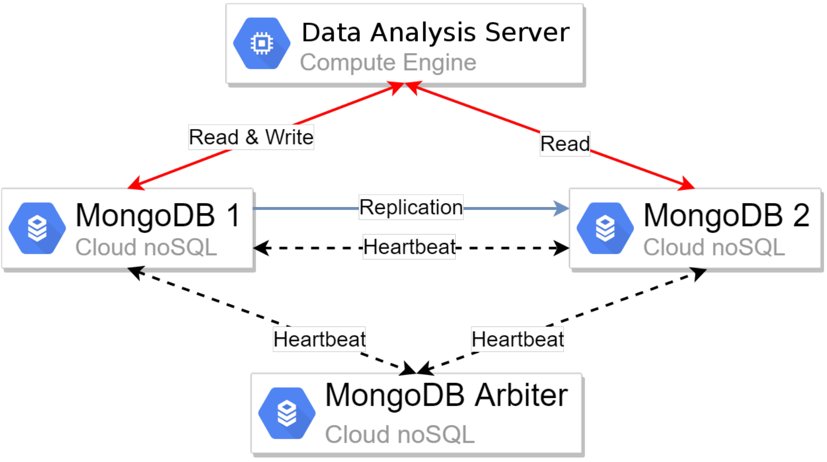


Figure 4.4: Storage architecture.

account on the PmA platform. A portion of edge computing is implemented within stations that have to pre-process the data acquired and guarantee that the whole collection process is GDPR compliant (further details are presented in section 4.3.3). Finally, the acquired data are encapsulated within JSON objects, together with some other information such as station identifiers (*e.g.* ID), location, and UNIX timestamp, before sending the data to the PmA Platform. Once this is done, no trace of the data collected remains on the station. The acquired data remains on board only for the time that is necessary to complete the scan cycle, which is set to a customisable default of 15 seconds.

4.3.2 PmA Platform

The PmA platform was designed and developed with a distributed architecture, so as to facilitate the scalability of the whole system. In fact, the system is distributed across four virtual machines. Three of these machines are dedicated to data storage only and are configured as a particular configuration of MongoDB named “replica set”. In this configuration there are two main roles: the *arbiter* and the *primary/secondary*. Figure 4.4 describes the architecture of the storage system. The arbiter machine is responsible to manage the coexistence between the “primary” and “secondary” databases, so as to prevent any disruptions. It decides which of the other two machines is selected as the “primary” one. At the same time the “secondary” copies and stores all the data present in the “primary”.

The operation is very simple; the arbiter continuously checks the availability of the “primary” and, as long as it is available, it stores the data in it. This is done through pings, or “heartbeats”. If the “primary” fails, the arbiter chooses a new “primary” from the “secondary” available. Once a “secondary” becomes “primary” the data is stored in the latter. In this way, the administrator has the time to manage the problem on the corrupted database.

The replica set configuration allows the system to be robust, due to the implementation of automatic data backup and the detection and correction of all errors in the primary. In case of failure, there is the possibility to immediately restore the system and make the existing data available again. Another machine implements front-end and backend logic and manages the communication flow with the primary database to send and receive processed data. Finally, it makes them available to the user through a management dashboard, where is possible to perform different operations.

The system described is entirely hosted in the Google Cloud Platform, using several services such as Compute Engine for the virtual machine, Cloud DNS in order to resolver addresses, Stackdriver for logging, and Cloud Functions to do batch operations during the night. The cloud functions are particularly useful to compute asynchronously the system post-processing metrics. A more detailed explanation is provided in the following.

4.3.3 Privacy aspects

In the last decade, the topic of privacy has become a prominent issue in any system that collects and processes data, particularly user-related information. In Europe, the General Data Protection Regulation (G.D.P.R. n. 2016/679[2]) defines the data content that can be exploited to identify an individual as “Personal Identification Information (PII)”, providing a specific indication of which data should be considered as personal information.

In the Internet of Things arena (but not limited to it) MAC addresses and IP

[2] https://www.gdpr.eu

addresses have always been a problem for users privacy, due to a lack of regulations in this regard. However, after the enactment of GDPR, both IP and MAC addresses must be treated as PIIs (art. 4 of GDPR regulation). In order to understand if there are privacy issues due to the data acquisition, the system defined by the European legislator was used which is based on the assessment of the risk (for the rights and freedoms of natural persons) deriving from the specific processing of personal data.

Security measures must be implemented by the person who has been tasked with the role of “data controller”. Data controllers must define security measures on the basis of a risk assessment. Furthermore, the data controllers must always provide maximum transparency on the purposes and methods of processing personal data. They must allow the data subject to control data processing, by making the rights provided for in the regulation easily and effectively manageable. Therefore a careful analysis of the specific reference context is necessary in order to respect all the phases of the data treatment. In this case, a preliminary assessment is needed on the type of data processed, to select only the data necessary to pursue the purpose of the processing. Unnecessary data must be deleted and data for which it is not necessary to maintain a connection with the identity of the persons must be made anonymous. Instead, the information that may be needed to reconnect to the persons concerned must be pseudonymised. In this way, the data controller can reduce risks and apply appropriate countermeasures.

For the reasons mentioned above and to guarantee people anonymity, in the proposed system all the collected MAC addresses are pseudonymised. All PmA stations send to the PmA platform a dummy identifier generated by applying sha256 encryption to all the collected MAC addresses. As result, the original MAC addresses are stored in neither stations nor the cloud and cannot be recovered performing the reverse operation.

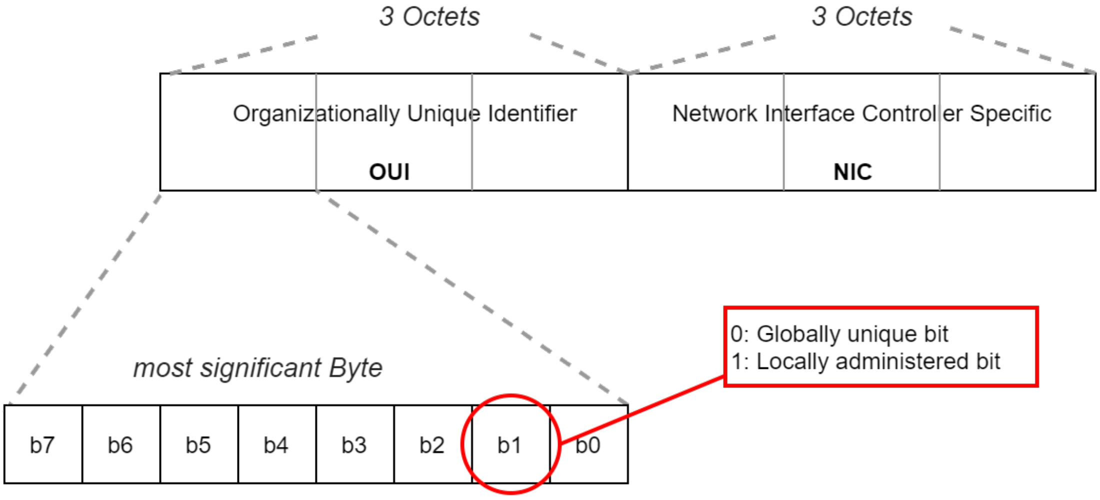

3 Octets
3 Octets
Organizationally Unique Identifier
OUI
Network Interface Controller Specific
NIC
most significant Byte
b7
b6
b5
b4
b3
b2
b1
b0
0: Globally unique bit
1: Locally administered bit

Figure 4.5: MAC address detail.

4.4 Data Analysis

This section show how data analysis was carried out on the data extracted from the "Probe Request" collected by the PmA stations, which are located in urban areas. Probe request frames are composed of several fields, but only a few are used to extract information. The most interesting field is the "source", because it contains the MAC address of the device that has sent the probe.

Figure 4.5 shows key details about the MAC address which includes 6 bytes in length, uniquely assigned by the manufacturer to each network card. The first three octets, referred to as OUI (Organisation Unique Identifier), are directly assigned by IEEE to the individual manufacturers of devices compatible with the Ethernet standard; the following three octets, called NIC (Network Interface Controller), are assigned by the device manufacturer, to ensure the addresses uniqueness. Looking at the second least significant bit of the first octet of the MAC address (as shown in red in Figure 4.5), this could be administered either universally (setting it to zero) or locally by the end-devices (when set to one). A universally administered MAC address is globally unique; this is not the case with a locally administered MAC address. This latter option is used to protect the users' privacy, for instance by periodically randomising the MAC address, which allows setting fake MACs to make it more difficult to track devices.

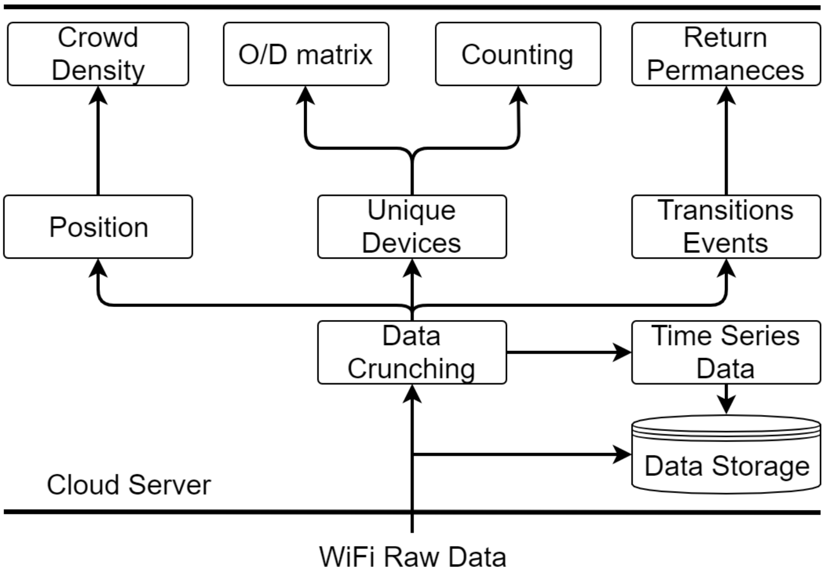


Figure 4.6: Data Flow.

User's tracking in relation to privacy has gained significant importance, to the point that the IEEE 802.11 working group has created a Topic Interest Group (TIG) on Randomised and Changing MAC addresses (RCM)[3]. This TIG is also focusing on the other issues of MAC addresses randomisation, such as network analytics and troubleshooting, network performance, device manufacturer identification, MAC-based Billing and Access Control, and the need for a standard covering the whole randomisation process.

One of the issues created by MAC randomisation is that it makes it hard to perform necessary data analysis tasks, such as device counting and localisation. In this work, all the probe request where the MAC address was locally administered were discarded.

Having clarified data collection, now it is possible to get into the overall data processing flow, as shown in Figure 4.6, whereby each layer of the diagram adds value to the data. Firstly, the raw data is processed by the *Data Crunching* module, which is responsible for creating the time series (list of data points sorted in time) and saving them on the data storage. The time series are then processed by position, device, and transitions/events, to obtain the output metrics.

[3] https://mentor.ieee.org/802.11/documents?is_dcn=DCN%2C%20Title%2C%20Author%20or%20Affiliation&is_group=0rcm

We now explain each module in greater detail. The localisation module computes the coordinates for each MAC address, within specific time ranges. The unique device module analyses all the MAC addresses received, returning a list of unique MAC addresses, which is then used by the other modules. The transitions events module computes all the presence transitions, allowing the Return Permanences module to compute the specific metrics. By this process it is possible to identify all the unique devices seen during the whole monitoring process, allowing to obtain the count of people. It is also possible to identify which users are stationary, which ones are returning to previous locations, and the duration of each event/transition (further details can be found in the following subsections).

The output metrics provided by the system can be divided into two categories:

- Real-Time metrics:
  - *Counting.* Number of devices detected by every single station within a certain *counting time range*, which can be chosen among different values (*e.g.* last hour, last day, custom range);
  - *Position.* Obtained via a trilateration method based on Friis' formula, using the power of a received signal and the transmission frequency.
- Post-processing metrics:
  - *Site returns.* This indicates after how long a device returns to the same place. It shows the number of devices that have come back after 5, 10, 30, 60, 120, 240, and 480 min, respectively;
  - *Site permanences.* This is similar to the return metric; it indicates for how long a device has been seen at a given place; the considered intervals are the same as the ones used by the return metric;
  - *O/D Matrix.* This shows how people have moved within a given PmA cluster; data is provided with a minimum interval of one day;
  - *Crowd density.* Starting from the single person localisation within the monitored perimeter, people's density is shown using heat maps.

In the following subsections, the metrics mentioned above are described in detail.

4.4.1 Counting

The system provides real-time information about the number of people who are present nearby the sensing stations. It is possible to see the evolution over time of this information, considering different aggregation times, such as 1, 10, 30 minutes, 1 hour or 1 day. Furthermore, once the count is received by the PmA platform, the percentage increase or decrease of the crowd with respect to the previous data is calculated. However, this metric is affected by an intrinsic error due to some people carrying more than or less than a single device on them. Another source of error arises from MAC address spoofing, which is performed by some smartphone models. To minimise errors, these types of MAC addresses were filtered out.

4.4.2 Site Returns

The system is able to recognise people coming back to the same station. Specifically, it can show a chart bar with different bins of time, such as 5 to 10 minutes, 10 to 30 minutes, 30 to 60 minutes, 60 to 120 minutes, 120 to 240 minutes, 240 to 480 minutes, and greater than 480 minutes, for any given day. The algorithm requires the setting of probe requests collected during a given *day*. The shape of the set operation, for each station i is described in the formula (4.1) below:

$$Probe_{day,i} = \{mac, timestamp_n\} \tag{4.1}$$

Collecting this information is useful to derive interesting information in regard to people habits and mobility nearby the stations. Taken in input the set of probe requests acquired by a given station ”i”, on a specific day ”day”, it is reordered according to the acquisition time in descending order. The consecutive arrival times of the different probes are then compared, obtaining deltas. If the delta is greater than five minutes, the corresponding bin is increased by 1. The five minute threshold was empirically chosen to ensure that the user was not really around the sensor. It was therefore assumed that all consecutive acquisitions received at a distance of less than five minutes are considered as a presence of the user in the monitoring area. In Algorithm 2 is reported the pseudo-code that describes this process starting from the raw data.

Algorithm 2 Site Returns Algorithm.

```
Input: $Probe_{day,i}$ descending order
  for all $timestamp_n \in Probe_{day,i}$ do
      $\Delta_t = timestamp_n - timestamp_{n-1}$
      if $\Delta_t < 5$ minutes then
          continue
      else if $5 <= \Delta_t <= 10$ then
          $bin5to10++$
      else if $10 < \Delta_t <= 30$ then
          $bin10to30++$
      else if $30 < \Delta_t <= 60$ then
          $bin30to60++$
      else if $60 < \Delta_t <= 120$ then
          $bin60to120++$
      else if $120 < \Delta_t <= 240$ then
          $bin120to240++$
      else if $240 < \Delta_t <= 480$ then
          $bin240to480++$
      else if $480 < \Delta_t$ then
          $bin480more++$
```

4.4.3 Site Permanence

The basic idea of this metric is complementary to the previous metric. How much time people remain in close proximity to specific stations is analysed. The algorithm 3 shows the pseudo-code used to extract this information. All probe requests captured by sensor i on the specific *day* are considered. For each of them the difference in the arrival time is calculated with respect to the previous one, obtaining Δ_t. If this difference is less than five minutes, different counters are updated; the first one is *count* which takes into account how many consecutive packets have been acquired, while the second is t_{perm} which takes into account the total permanence time. As also explained in the previous subsection 4.4.2, the five minute threshold was empirically chosen as maximum time between two consecutive probes in order to consider the sending device still into the monitored area by the sensor. If the delta should be greater than five minutes, this means that the user has moved away from the area under examination, therefore the total residence time is checked and the devices' number in that specific time bin is increased by one.

Algorithm 3 Site Permanence Algorithm.

```
Input: $Probe_{day,i}$ descending order
Input: $count = 0$
Input: $t_{perm} = 0$
  for all $timestamp_n \in Probe_{day,i}$ do
    if $timestamp_n == timestamp_0$ then
      $\Delta_t = timestamp_n$
    else
      $\Delta_t = timestamp_n - timestamp_{n-1}$
    if $\Delta_t < 5\ minutes$ then
      $count{+}{+}$
      $t_{perm} = t_{perm} + \Delta_t$
    else
      if $count > 3$ then
        if $5 <= t_{perm} <= 10$ then
          $bin_{5to10}{+}{+}$
        else if $10 < t_{perm} <= 30$ then
          $bin_{10to30}{+}{+}$
        else if $30 < t_{perm} <= 60$ then
          $bin_{30to60}{+}{+}$
        else if $60 < t_{perm} <= 120$ then
          $bin_{60to120}{+}{+}$
        else if $120 < t_{perm} <= 240$ then
          $bin_{120to240}{+}{+}$
        else if $240 < t_{perm} <= 480$ then
          $bin_{240to480}{+}{+}$
        else if $480 < t_{perm}$ then
          $bin_{480more}{+}{+}$
      $count = 0$
      $perm = 0$
```

4.4.4 Origin/Destination Matrix

Origin and destination are crucial for the planning of transit routes and stop locations. These are usually collected by Telecom Operators or are simply based on travellers surveys. However, these approaches come with significant drawbacks. The first one is that Telecom Operator data are very expensive, furthermore, they are not broadly representative because of their strong dependence on the market share held by the single operator. On the other hand, survey data are known to be susceptible to statistical weakness, bias, and other general errors [94].

The PmA system can recognise the presence of devices already seen by other stations. Thanks to this information, it is possible to estimate an Origin-Destination (O/D) matrix, associated with the stations deployed around the monitored zone. For this metric, in the PmA platform, the standard observation interval is a whole day. Yet, it is possible to extend the interval to several days.

In the matrix, each station identifies a specific zone. In order to compute the O/D matrix, the system looks for any MAC addresses that were seen by at least two stations during the selected time window. After searching for this information, the list of MAC addresses is sorted and transitions from one zone to another are found and added to the matrix box. To clarify, if the system detects that a user moves from station A to station B, the number of movements from A to B is updated. The process is repeated for all other stations, organised in clusters. This helps to understand how people move and identify the most popular routes.

4.4.5 Crowd Density

The PmA platform can create heat maps directly correlated to people's density, within specific monitored areas. Although several papers have dealt with indoor/outdoor localisation based on tracking via Wi-Fi [95, 96, 97], in this work the focus is on RSSI-based and Time Difference of Arrival techniques. Selecting a time range in which to compute this information, PmA exploits some probe request packet fields to estimate the position of the devices, as further explained in the following two sub-sections that deal with device localisation.

RSSI-based localisation

The first proposed algorithm belongs to the family of range-based algorithms, which is based on the RSSI value contained within the probe request. Major efforts have been done by scientists to understand and improve these techniques [98, 99, 100], for which reference is made to the Stefan Knauth's work [101].

In PmA, a range-based algorithm derived by the Friis' transmission formula [102] is used, using the frequencies of Wi-Fi communication, as reported in equation (4.2), where it is figured out how to calculate the distance d between a PmA station with known coordinates (anchor) and the target device (whose position is unknown). The use the following formula, assuming omnidirectional antennas:

$$d = \frac{\lambda}{4\pi\sqrt{\frac{P_{rx}}{P_{tx}}}} \tag{4.2}$$

where λ is the wavelength computed at 2,4 Ghz and P_{rx} and P_{tx} are the power of receiver (PmA Station) and transmitter (*e.g.* smartphone), respectively. Given a set of MAC address seen by the sniffing stations within a scanning window $\mathcal{S}_w$, it is possible to define an MAC address group $\mathcal{M}_t$ seen by multiple stations during a time-slot t in the scanning window S_w. Each element of $\mathcal{M}_t$ is represented by:

$$m_{t,i} = \{mac_i, s_{lat,i}, s_{lon,i}, P_{rx,i,t}\} \tag{4.3}$$

where $s_{lat,i}$ and $s_{lon,i}$ are, respectively, the latitude and longitude of the station i that has "seen" the MAC address mac_i. Finally $P_{rx,i,t}$ is the power contained within the Wi-Fi probe request. From this information, it is possible to compute trilateration in order to derive an approximation of the Wi-Fi device's position by means of algorithm 4.

Algorithm 4 PmA Derive Positions algorithm.

```
for all $t \in \mathcal{S}_w$ do
    for all $mac \in \mathcal{M}_t$ do
        $itx_{t,mac} = Intersections(mac)$
        if $itx_{t,mac} = Null$ then
            Set $P_{t,mac}$ with Friis's based positioning
        else
            Set $P_{t,mac} = CoG(itx_{t,mac})$
```

The algorithm calculates the centre of gravity (CoG) of the polygon resulting from the intersections of the circumferences $itx_{t,mac}$. The circumference radius is equal to the distance d computed in (4.2), considering P_{rx} as the RSSI contained into the probe request packet and P_{tx} as the mean power for a common Wi-

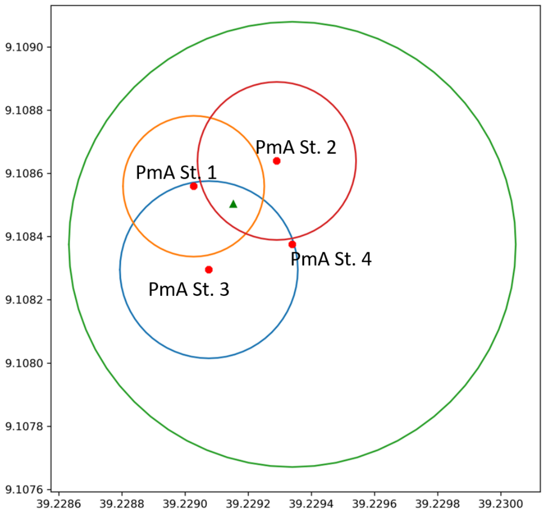


Figure 4.7: Target position estimated by RSSI-based algorithm.

Fi antenna[4]. If the target device is seen only by one station, there is only one circumference, and it is not possible to find intersection points. In this case, the present algorithm chooses a random point that lies on the circumference with radius given by Friis's formula, using the power contained in the probe request.

Figure 4.7 shows the results of RSSI-based localisation in the controlled scenario, whereby it is possible to identify the intersections of circumferences. Thanks to the position estimation obtained from the algorithm, it is possible to obtain the crowd density, as shown in section 4.5.2, by means of a heat map.

TDOA-based localisation

In this section how to use the Wi-Fi probe request frame to derive the device's position is explained. The basic idea is to collect the Time of Arrival (ToA) of the probe request management frame from each station and, then, consider the packets having the same sequence number within a short scanning window (*e.g.* 1 second). Subsequently, this information is used to solve the following system of equations.

Let us define the target position with $P_t(x, y)$ and the anchor-stations positions as $P_a(x_a, y_a)$, $P_b(x_b, y_b)$ and $P_r(x_r, y_r)$. Furthermore, the time taken by the signal

[4] https://android.googlesource.com/platform/frameworks/base/+/master/core/res/res/xml/power_profile.xml

emitted by the target to reach the stations as:

$$\begin{cases} T_a = \frac{1}{c}(\sqrt{(x-x_a)^2+(y-y_a)^2} \\ T_b = \frac{1}{c}(\sqrt{(x-x_b)^2+(y-y_b)^2} \\ T_r = \frac{1}{c}(\sqrt{(x-x_c)^2+(y-y_c)^2} \end{cases} \tag{4.4}$$

Where c is the speed of light. Let us take P_r as a reference anchor. Accordingly, it is possible to define the differences between the previous arrival times, as:

$$\begin{cases} \tau_a = T_a - T_r = \frac{1}{c}(\sqrt{(x-x_a)^2+(y-y_a)^2} - \sqrt{x^2+y^2}) \\ \tau_b = T_b - T_r = \frac{1}{c}(\sqrt{(x-x_b)^2+(y-y_b)^2} - \sqrt{x^2+y^2}) \end{cases} \tag{4.5}$$

Without loss of generality, the general and system equations could be rewritten as:

$$\begin{cases} (x-x_r)^2+(y-y_r)^2 = d_r^2 \\ (x-x_a)^2+(y-y_a)^2 = (d_r+l_{a,r})^2 \\ (x-x_b)^2+(y-y_b)^2 = (d_r+l_{b,r})^2 \\ \cdots \\ (x-x_n)^2+(y-y_n)^2 = (d_r+l_{n,r})^2 \end{cases} \tag{4.6}$$

Where d_i is the distance between the target point P_t and the $i-th$ anchor point, $l_{i,r}$ is the TDOA range estimation. To improve readability let us substitute:

$$(x_i - x_r) = \overline{x_i} \quad and \quad (x - x_r) = \overline{x} \tag{4.7}$$

This form of equations 4.6 is quite hard to understand for a calculator and is rather inefficient. Therefore, in the implementation, the *Least Squares Method* was used to simplify and solve the equations. Finally, from 4.6 subtracting the first one at the other equations and putting them in matrix form, the system of equations can be rewritten as follow:

$$2\begin{bmatrix}\overline{x_a} & \overline{y_a}\\ \overline{x_b} & \overline{y_b}\\ \cdots & \cdots\\ \overline{x_n} & \overline{y_n}\end{bmatrix}\begin{bmatrix}\overline{x}\\ \overline{y}\end{bmatrix}=\begin{bmatrix}\mu_a-l^2_{a,r}\\ \mu_b-l^2_{b,r}\\ \cdots\\ \mu_n-l^2_{n,r}\end{bmatrix}+d_r\begin{bmatrix}-l_{a,r}\\ -l_{b,r}\\ \cdots\\ -l_n,r\end{bmatrix} \tag{4.8}$$

Where $\mu_i = \|P_i\|_2^2 = x_i^2 + y_i^2$. Given the following substitutions:

$$\underline{A}=\begin{bmatrix}\overline{x_a} & \overline{y_a}\\ \overline{x_b} & \overline{y_b}\\ \cdots & \cdots\\ \overline{x_n} & \overline{y_n}\end{bmatrix},\quad \underline{X}=\begin{bmatrix}\overline{x}\\ \overline{y}\end{bmatrix},\quad \underline{\Phi}=\begin{bmatrix}\mu_a-l^2_{a,r}\\ \mu_b-l^2_{b,r}\\ \cdots\\ \mu_n-l^2_{n,r}\end{bmatrix},\quad \underline{\Lambda}=\begin{bmatrix}-l_{a,r}\\ -l_{b,r}\\ \cdots\\ -l_n,r\end{bmatrix} \tag{4.9}$$

The matrix equation can be described as:

$$2\underline{A}\,\underline{X}=\underline{\Phi}+d_r\underline{\Lambda} \tag{4.10}$$

The equation could be solved by means the *Least Squares Method* [103] and its solution is:

$$\underline{X}=\frac{1}{2}(\underline{A}^t\underline{A})^{-1}\underline{A}^t(\underline{\Phi}+d_r\underline{\Lambda}) \tag{4.11}$$

The previous equation contains the parameter d_r and constitutes a non-linear expression. Therefore, solving the equation 4.11 leads to the final solution and identifying the target.

In Figure 4.8 it is shown how the algorithm computes the solution of equations 4.11, using one of the four PmA stations as a reference anchor. In particular, a simulation in a Cartesian plane was done, where all parameters (*i.e.* time of arrival and distances) had already been computed.

4.5 Experimental analysis

The experimental activity has been conducted in both a controlled scenario (to evaluate the performance of the positioning algorithm) and in real scenarios

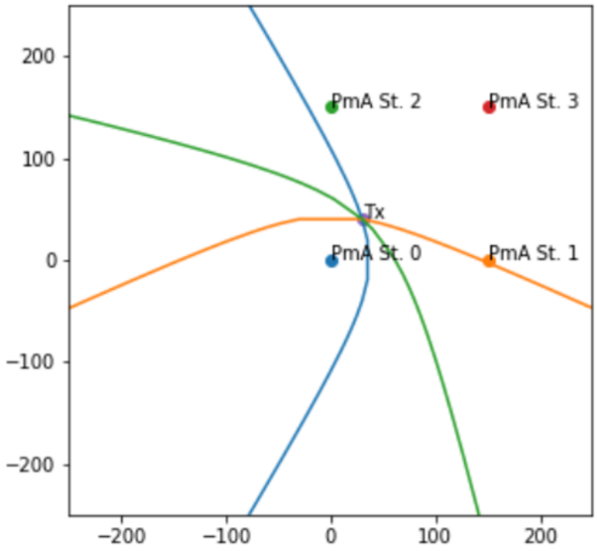


Figure 4.8: Target position estimated by TDOA-based algorithm.

through different pilot studies (to evaluate the performance of the other metrics). In the following, as first it is presented the setting and, then, the analysis of the results.

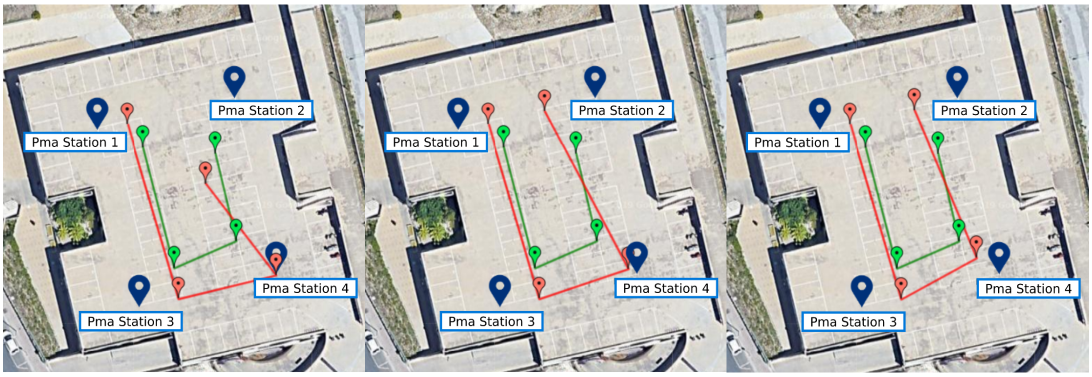


Figure 4.9: Ground truth trajectory (green) and trajectory (red) computed with RSSI-based algorithm. *Left:* 40 seconds of time-aggregation; *Center:* 80 seconds of time-aggregation; *Right:* 120 seconds of time-aggregation. The experiment was done in the Faculty of Engineering, University of Cagliari.

To test and validate the localisation algorithms, an outdoor empty open space has been selected, namely a non-utilised parking area of the Engineering Faculty of the University of Cagliari. This area has been selected due to the absence of obstacles and objects, which could have otherwise interfered with the stations. As shown in 4.10, the experiment has been performed using four PmA stations (black dots). In Figure 4.11 it is possible to see the stations used for testing in

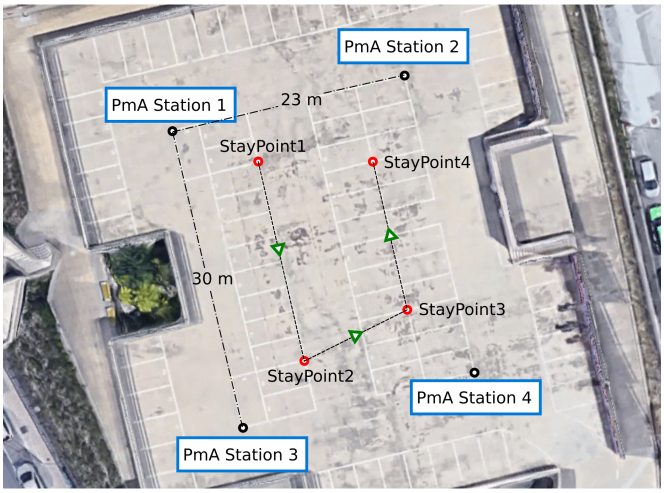


Figure 4.10: Map of the controlled scenario area.

the controlled scenario. These were placed at the corners of a rectangle with a perimeter of 107 meters and an area of 710 square meters.

4.5.1 Experiments setup

Controlled scenario

Once the stations were in place, a path with an Android smartphone that was collecting the position in order to have a ground truth useful for performance analysis was followed. In Figure 4.10 in red are marked the stay-points of the followed path, spending 3 minutes for each point, to be sure that the station would collect enough data for the experiment. The basic idea of the experiment was to try and understand the level of reliability of the two methods (RSSI-based and TDOA-based) applied to a pedestrian mobility scenario.

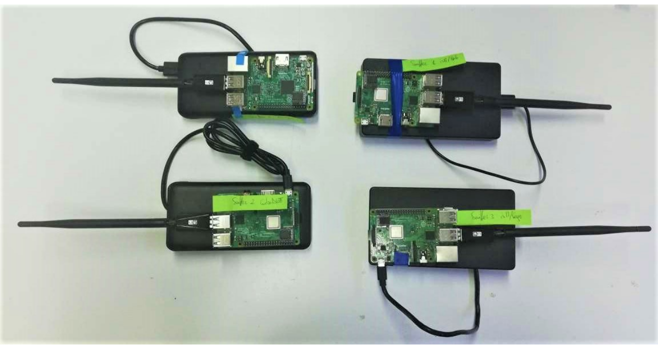


Figure 4.11: PmA Stations used for tests in the controlled scenario.

Pilot studies in real-world scenarios

Different experiments in real-world cases were also performed. The data were acquired on three different scenarios: in the city centre of Turin (Italy); during the International Truffle Festival in the city of Alba (Italy); and at the Engineering Faculty of the University of Cagliari (Italy). Each of these installations was created for a specific use case. The Turin centre experiment was characterised by the following features:

- Objective: device counting near a roundabout, monitoring device on-site returns and on-site permanency, flow of vehicular traffic with O/D matrix.
- Installed devices: the devices were installed in a roundabout with 6 confluent arteries. A sniffing station was installed for each road, with a distance of about 20 meters from the entrance to the roundabout.

The Alba International Truffle Festival experiment had the following features:

- Objective: crowd density in the historic centre;
- Installed devices: 5 stations were installed in points of interest, identified in the historic centre of Alba. Each station was installed near the road to facilitate data acquisition. The purpose of this installation was to understand how many people visited the points of interest during the truffle fair.

The Engineering Faculty experiment was characterised by the following features:

- Objective: people counting, crowd density.
- Installed devices: 8 sniffers have been installed, covering the area of the park in the Engineering faculty, to monitor overcrowding during the day and count people.

The device's configuration used in these experiments is depicted in Figure 4.3.

Table 4.3: Error evaluation in meters.

	Aggregation [seconds]				
	40	**60**	**80**	**100**	**120**
Point 1	4.5	4.5	4.5	4.5	4.5
Point 2	5.1	5.1	5.1	5.1	5.1
Point 3	7.0	7.0	7.3	7.0	7.7
Point 4	10.3	8.3	8.1	8.3	7.2
RMSE	14.2	12.8	12.9	12.8	12.5

4.5.2 Experimental results

Controlled scenario

The experiments were conducted following the setting explained in section 4.5.1 and the results are summarised in Table 4.3. After data collection, in-depth data processing was performed.

Initially, all the probe requests captured during a specific time interval determined by the aggregation parameter were selected. This parameter is used to increase or decrease the time interval centred at each moment in which the target remained stationary in the stay points (red marker in Figure 4.10). Then all the probe requests inside a time interval are taken into account for the estimation of the position. To compare localisation errors at the different stay points, the Root Mean Square Error (RMSE) was used as a gauge.

Evaluating the results showed that in Point 1 and Point 2, the error is fixed for each aggregation interval. This behaviour is due to the usage of the median to find out the average power from the probe received. The outliers points have less weight so the median value is stable across the different aggregations. A different situation appears for Point 3 and Point 4, where the median error changes somewhat more unpredictably. This is due to scattering and multi-path effects that are most probably responsible for interference in the signal propagation.

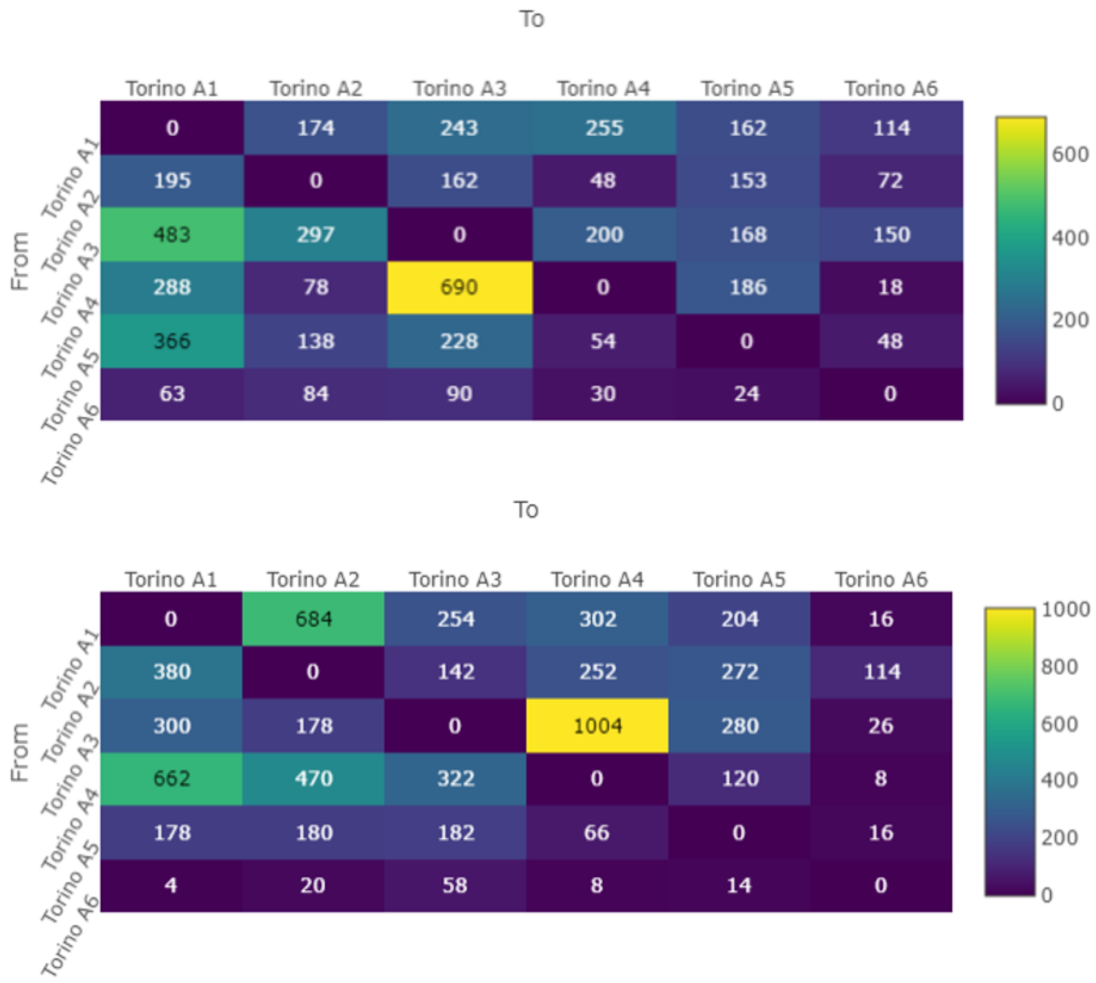
In general, it is possible to notice that, by increasing the time observation interval, RMSE decreases. Obviously, this is true only if the target is staying still at the same point for an amount of time comparable to the window time interval.

At first, the TDOA algorithm was performed through simulation, leading to very promising results. Then the same experiment in a pilot study was performed, employing the RSSI algorithm. Unfortunately, the experiments have shown that by means of the PmA stations equipped with this low-cost hardware, is not possible to obtain satisfactory results. As a matter of fact, the weakness in this type of approach is represented by the time measurements accuracy and precision, which requires hardware with extremely precise time resolutions. Thus, it is crucial to employ higher-spec hardware as anchor device, and pursue strong time synchronisation among the different anchors, as preconditions for an effective TDOA-based algorithm in a real-world scenario with short distances among anchors.

Further problems, in relation to the event chosen as the trigger in the time-counting systems or the latency between the different process layers (eg coding, synchronisation, etc.) were identified. Moreover, multipath and NLOS problems may occur between transmitter and receiver. Finally, since the PmA stations are based on Raspberry Pi and Linux OS, the OS process scheduling policy could not be a precise time-acquisition process.

Without extremely accurate synchronisation at the moment it is not possible to reach an accuracy in the nanosecond range, which is the key limitation pinpointed in real-world settings. One possible solution is to implement the GPS or the Precision Time Protocol IEEE 1588 [5]. The adoption of the IEEE 1588 Protocol would indeed allow us to fix this important issue in the TDOA algorithm. At the moment, a raspberry version of Linux-ptp exists, which, however, is currently incompatible with the most recent versions of Raspbian OS, necessary to run the developed scripts to perform the other operations, such as probe requests collection and processing. Another obstacle for rapid implementation of this solution is that the IEEE-1588 protocol was designed to work on LAN networks, and not on WLAN networks, as it is in this case. However, this issue could be fixed by following the solution proposed in [104], which would require further investigation for a full understanding.

[5] https://www.nist.gov/el/intelligent-systems-division-73500/ieee-1588



From \ To	Torino A1	Torino A2	Torino A3	Torino A4	Torino A5	Torino A6
Torino A1	0	174	243	255	162	114
Torino A2	195	0	162	48	153	72
Torino A3	483	297	0	200	168	150
Torino A4	288	78	690	0	186	18
Torino A5	366	138	228	54	0	48
Torino A6	63	84	90	30	24	0

From \ To	Torino A1	Torino A2	Torino A3	Torino A4	Torino A5	Torino A6
Torino A1	0	684	254	302	204	16
Torino A2	380	0	142	252	272	114
Torino A3	300	178	0	1004	280	26
Torino A4	662	470	322	0	120	8
Torino A5	178	180	182	66	0	16
Torino A6	4	20	58	8	14	0

Figure 4.12: *Up*: O/D Matrix morning - scanning window *Down*: O/D Matrix evening - scanning window.

Pilot studies in real-world scenarios

Below the data collected from the various stations were analysed. In particular, in this section, the data relating to a cluster of stations in Turin (6 stations), a cluster deployed in Alba (5 stations), and a cluster deployed at the Faculty of Engineering, University of Cagliari (8 stations) are presented.

Turin. In the pilot in Turin, the most interesting metrics to be analysed are those related to the site returns and traffic flow. The first interesting point here is provided by the return devices (Figure 4.13). Monitoring was carried out in two-time ranges, morning and afternoon respectively.

Given the time slot, a considerable number of people can be seen returning to the

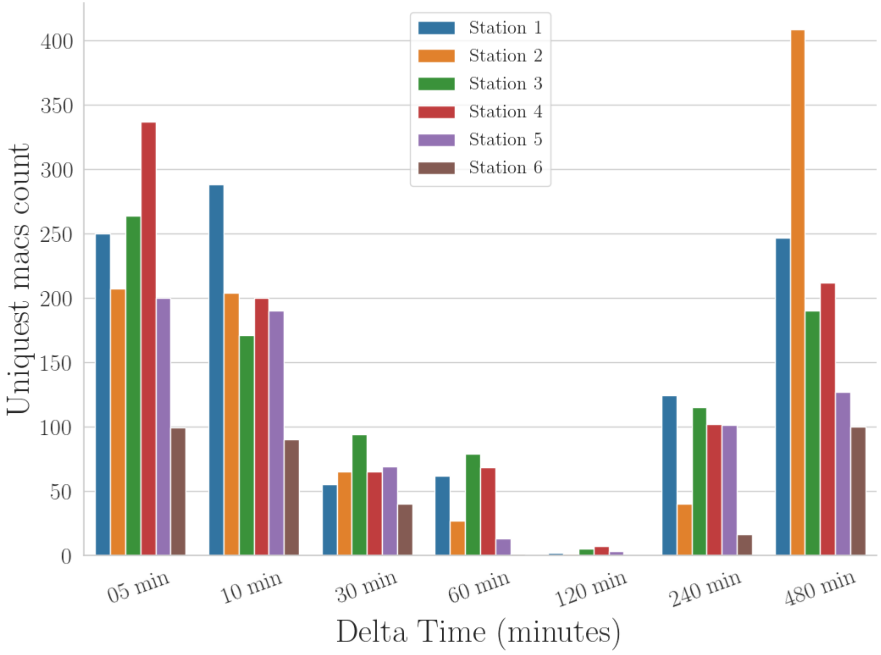


Figure 4.13: Number of returning devices (Turin).

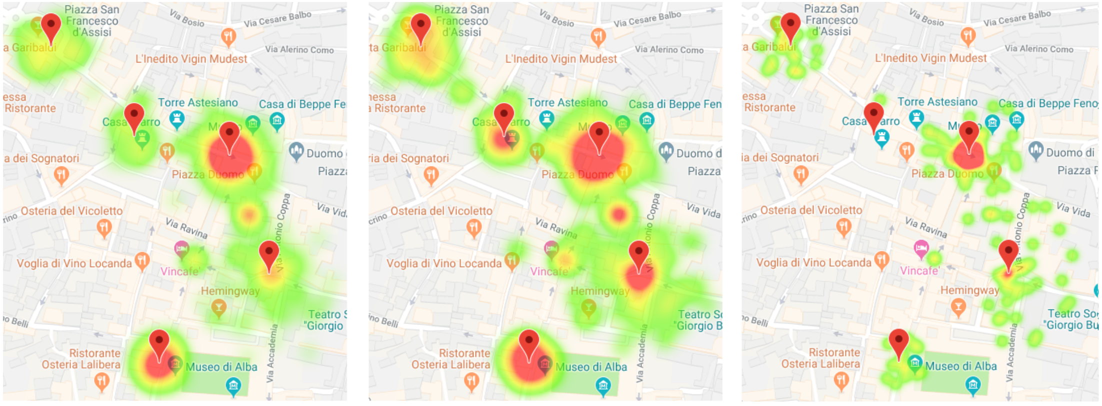


Figure 4.14: Crowd Density in the truffle fair in Alba.

stations after 480 minutes or more, this is attributable to people returning after a working day. By analysing the O/D matrix (see Figure 4.12) it is possible to see how the busiest exits and entrances of the roundabout during two intervals of time, morning and afternoon. As shown on the higher part of Figure 4.12, during the morning most of the devices have entered the roundabout near the Torino_A4 station coming out near the Torino_A3 station. On the contrary, in the afternoon (lower part of Figure 4.12), exactly the opposite happened.

Alba. Regarding the Alba pilot, the most interesting metric to analyse relates to crowd density. The installation was performed during the last two days of the truffle festival. In Figure 4.14 is shown the crowd density for the last two days of the festival and the following day when the festival is over. The first two maps show how the crowding was similar during the festival days. The most crowded place was the cathedral square where the main events took place. The third map shows the day after the festival. It can be seen that the density of people near the stations drops significantly, indicating that the flow of tourists has decreased considerably compared to the previous two days.

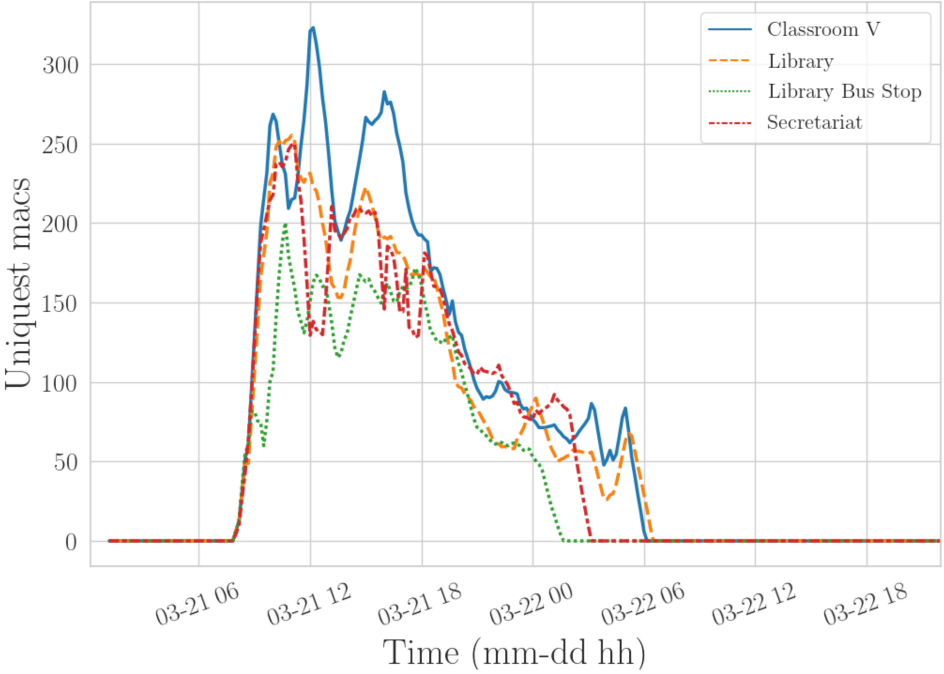


Figure 4.15: Uniquest MACs Distribution (Faculty of Engineering, University of Cagliari).

Faculty of Engineering, University of Cagliari. For the Faculty of Engineering, University of Cagliari pilot, the 8 stations were grouped into two clusters, containing 4 stations each. Cluster A contains stations installed close to the classrooms, library and secretariat. Cluster B contains the stations installed in the park. The most interesting metrics, in this case, are the people counting, site returns and site permanence for cluster A, and crowd density for both clusters.

For convenience, the analysis of Cluster A is the first displayed. In Figures 4.16 and 4.17, which respectively indicate site returns and site permanence, which immediately highlights a peak of about 120 unique MACs that returned after a 2-hour interval. This is also visible in Figure 4.15 where you see three peaks relating to classroom V. However in the chart there is a bias of about 50 MAC addresses due to Wi-Fi devices close the PmA cluster (*e.g.* Wi-Fi printers, APs and computer classroom notebooks, cars or pedestrian passing near the PmA cluster). The three peaks correspond to the times of the morning classes (10 am and 12 am) and the afternoon (4 pm). At 2:00 pm, fewer devices are detected due to the lunch break. In the next hour, however, the number of devices tends to increase, until it reaches the peak at 16.00. This justifies the number of site returns shown in Figure 4.16. Another interesting event shown in 4.17 relates to the time intervals of 5 and 10 minutes, showing how students may have taken short breaks. Another interesting data is given by the station installed in the secretariat, which reports the breaks of 5-10 minutes and the lunch breaks of 30-60 minutes taken by the employees.

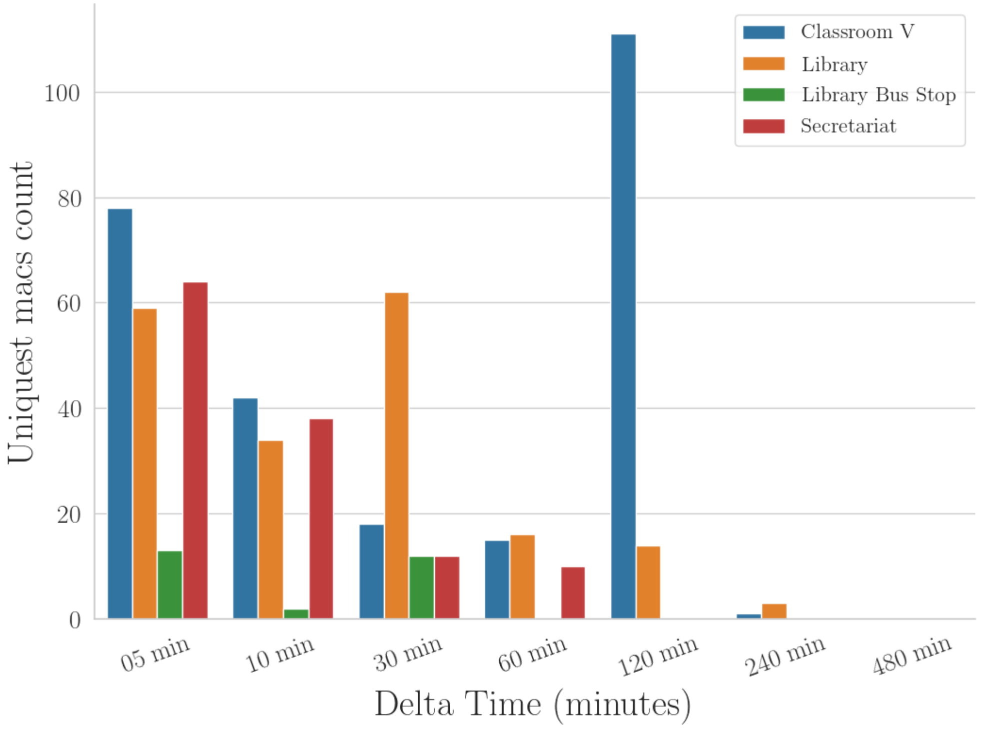


Figure 4.16: Number of returning devices at University of Cagliari.

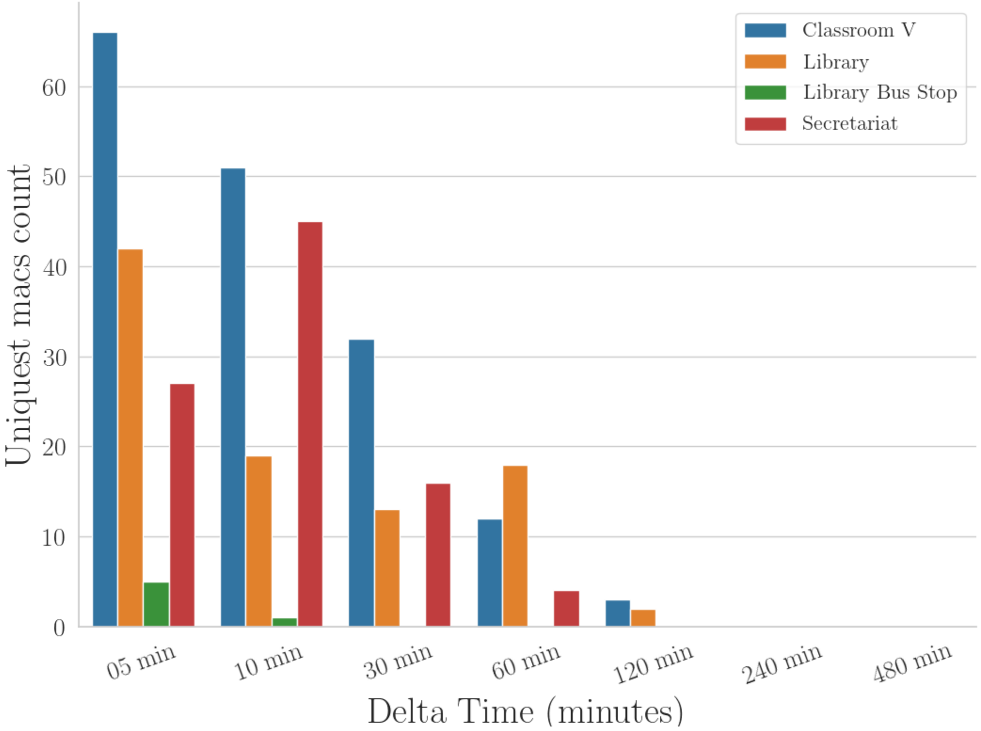


Figure 4.17: Number of stationary devices at University of Cagliari.

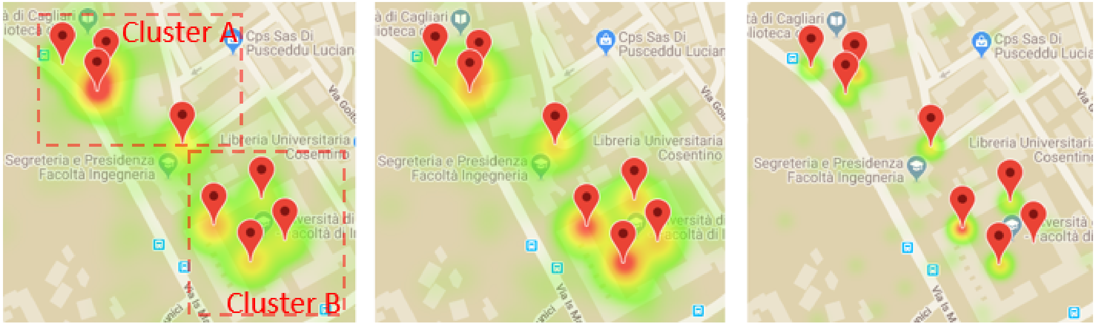


Figure 4.18: Crowd Density in the Faculty of Engineering, University of Cagliari.

Continuing the analysis, in the three snapshots shown in Figure 4.18, there is a view of the evolution of the crowding density, relative to the pilot campus. This is possible using the coordinates estimated by the proposed system. The first map indicates the crowded density at 11.00, the second at 14.00 and the third at 20.00. In the first map, the most crowded area seems to be related to cluster A. Indeed, here there are two of the largest classrooms on campus and the library. In the second test, made at lunchtime, the most crowded area is related to cluster B where the park is, which is very populated at that time. Finally, the third map shows a crowding much lower than that of the others. This is because at 20.00 the number of students and workers is very low.

4.6 Summary of key results

The aim of this work was to design and develop a complete framework for the acquisition and analysis of data relating to the mobility of people in urban areas through a passive strategy. Wanting to make the presented system as generalizable as possible, a low-cost device was designed, built and tested to monitor and acquire the necessary mobility data.

All the details of the system architecture were presented, consisting of clients (data acquisition sensors) and servers (where data management and analysis is performed). In order to obtain interesting insights from the data, different metrics are extracted during the data analysis, such as crowd density, the flows of people moving between the different areas, the O / D matrix (useful for transport service planning), returns and stay numbers. The solution was tested first under controlled conditions and then in three pilot implementations, proving to be extremely flexible and adaptable to different contexts.

This chapter responded to the lack of a complete and replicable framework for passively acquiring people mobility data and for data analysing in order to provide powerful metrics on people mobility behaviour.

Chapter 5

Mobility tracing through data analysis

All the resulting work presented in this chapter has been incorporated into a paper entitled "*Mobility Analysis during the 2020 Pandemic in a Touristic city: the Case of Cagliari*", accepted at 2021 IEEE IoT Vertical and Topical Summit for Tourism (IoT-VTST'21) [12].

5.1 Introduction

If we think of cities as the beating heart of human life, we could then see the city traffic as the blood flow, and the roads as the arteries. Following this metaphor, we can say that the traffic trace patterns not only tell us about the citizens' mobility patterns but also provide a good indication as to the efficiency of the road system. For these reasons, monitoring and analysing traffic flows, relative volumes and patterns are fundamental to making insights as to how people live in a city and how they react to different scenarios.

The COVID-19 pandemic has created a unique opportunity to make new insights about how people move in a city, discovering how patterns (not just volumes) have been changing in the various critical stages of the emergency. What is more, the lockdown restrictions have created a unique condition in which to analyse the critical mobility patterns, those that keep the city alive and our needs met. The opportunity was taken therefore to study the background traffic that has kept going a tourist city (*i.e.* Cagliari) during the pandemic. In this situation there was virtually no leisure traffic and there were no tourists around, leaving space only to the essential movements; *e.g.* emergency, medical, food delivery etc.

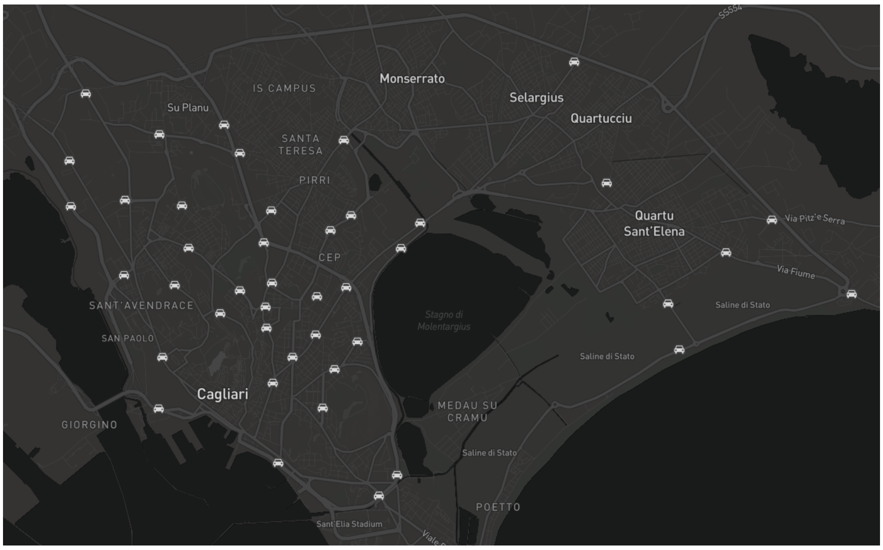


Figure 5.1: Traffic measurement stations in the city.

In this chapter, extensive mobility traces dataset is analysed, which is generated by 167 traffic sensors, organised in 98 measuring stations scattered around the city, and operating continuously since 2016. A schematic of the city and the sensing stations is included in Figure 5.1. Although many other mobility studies have been carried out since the start of the pandemic, as described in Section 5.2, a unique perspective is taken, focusing on traffic volume, patterns and tourist-related information.

The study is based entirely on the open data made available by the municipality[1]. However, significant effort has been required to pre-process the raw data, which is otherwise not directly usable due to issues arising from the data collection and transmission process. The data preparation method is presented in Section 5.3, going from the acquisition issues, through the detection of sustained anomalies, the selection of the acceptable data (filtering out non-reliable data) and then finally using machine learning to model the traffic.

[1] https://opendata.comune.cagliari.it/portale/it/st04_api_cloud.page

While some of the findings were expected (*e.g.* dramatic reduction of traffic volumes during lockdown), some interesting insights emerged (Section 5.4.1). First, it could be noted that, following an initial 76% traffic reduction on the first lockdown, subsequent restrictions have led to less sudden changes. While the absolute traffic volumes roughly followed the pandemic evolution, the weekly traffic patterns changed drastically over time, whereas the daily ones were more consistency. Traffic traces were also compared with official tourist-presence figures, which allowed us to detect the traffic stations most influenced by tourists' mobility.

5.2 Related work

The impact of COVID-19 on usual habits has been important and it has been studied at global level. Several studies have been done and different phenomena have been analysed around the world. In addition to the direct impact on health (for which most of the energy has been put to fight the spread of the infection), other important phenomena have occurred as consequences of the pandemic. Studies on the change in mobility have been carried out in different parts of the world, focusing mainly on the variation of road flows. In [105] Aletta et al. carried out an analysis of the effects on noise pollution given by the reduction in traffic caused by the mobility restrictions for the containment of COVID-19 in Rome, and reported a significant reduction in private vehicle trips in the city (-64.6 % during the lockdown). Yet, unlike the work presented in this chapter, their entire study was based on Floating Car Data[2].

In [106] Harantová et al. focused on comparing the range and speed of vehicles on a specific road section of the first-class road I/11, which is part of the European road E75. This was captured before and after the implementation of the mobility restrictive measures, to understand what effect the measures have had on the quality of traffic flow in the Slovak Republic. The difference with this work basically concerns the number of measurement points, 2 for the work mentioned and 98 for the work presented in the present study.

[2] https://en.wikipedia.org/wiki/Floating_car_data

Another reason why mobility patterns vary is due to social distancing. In [107] De Vos et al. discuss the implications on daily travel patterns, spotting how social distancing changes the way people cope with daily activities. They also show how social distancing has negatively affected subjective wellbeing and health, as people tend to travel less on public transport.

In [108] Wang et al. analysed mobility patterns at the state level, and found that changing patterns of mobility are not necessarily related to government policies and guidelines, but are more related to people's awareness of the pandemic. In their results, it is noted that it takes on average 14 days for the mobility models to adapt to the new situation. A similar situation is also observed in this work.

Although the reduction in traffic volumes has had an effect also on the reduction of road accidents (-41 % in the example of Greece), in [109] Katrakazas et al. found that in some states there was also an increase in incorrect behaviour (such as an increased average driving speed), an increase in the frequency of sudden acceleration and braking (up to 12% increase) and the use of mobile phones while driving (up to 42% increase). Another important consequence of the pandemic and of the restrictions necessary to contain the virus is the change in air quality. This aspect has not been explored, which can be referred to in several studies, in different worldwide states such as Brazil[110], India[111], Morocco[112], Kazakhstan[113], Spain[114, 115] just to mention a few.

In [116] Yang et al. analyse the population density and correlate the contagion data with the number of travellers and affirm that by monitoring traffic and air flows it is possible to mitigate the infections. On the other hand, in this study the total tourist traffic of the city within a specific area was compared, finding a strong correlation that indicates how from the traffic patterns it is possible to understand urban movements. Finally, in [117] Chen et al. examine urban traffic in Wuhan and present some metrics to manage the emergency.

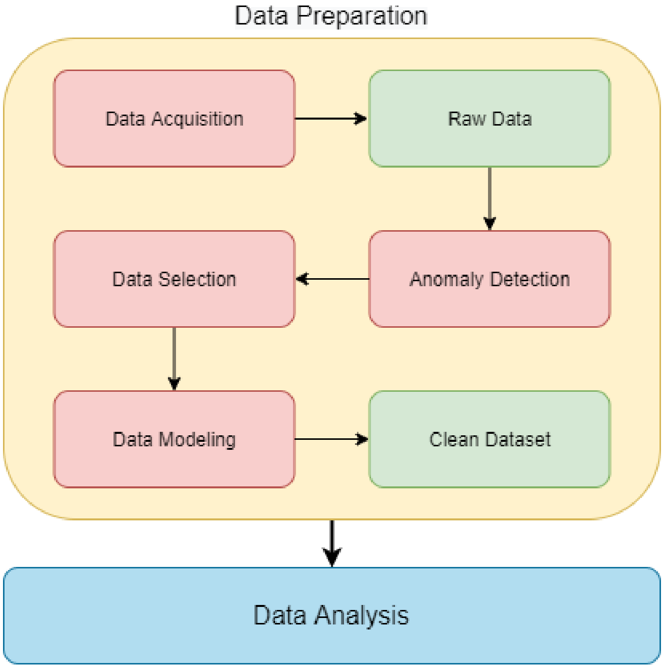


Figure 5.2: Data pipeline; from raw data to a clean dataset.

5.3 Data Science Pipeline

In this section, it is introduced the entire data science pipeline adopted herein, as depicted in Figure 5.2.

5.3.1 Data acquisition

In 2016 the Cagliari municipality installed sensors consisting of inductive loops located under the road surface to monitor the traffic, as exemplified in Figure 5.3. The data acquired is available as open data via API, and through the API all data comprised between 1st January 2016 and 31st December 2020 was downloaded.

5.3.2 Anomaly detection

After retrieving the data, an important step was to find the anomalous values. For this operation, two different levels of cleaning were used. A first selection was done choosing a rigid threshold of the maximum traffic value for each measurement station. This threshold takes into account the number of sensors (one per lane) that each station has. The larger the number of sensors a station has, the higher the number of lanes that the station monitors, as well as the greater the flow

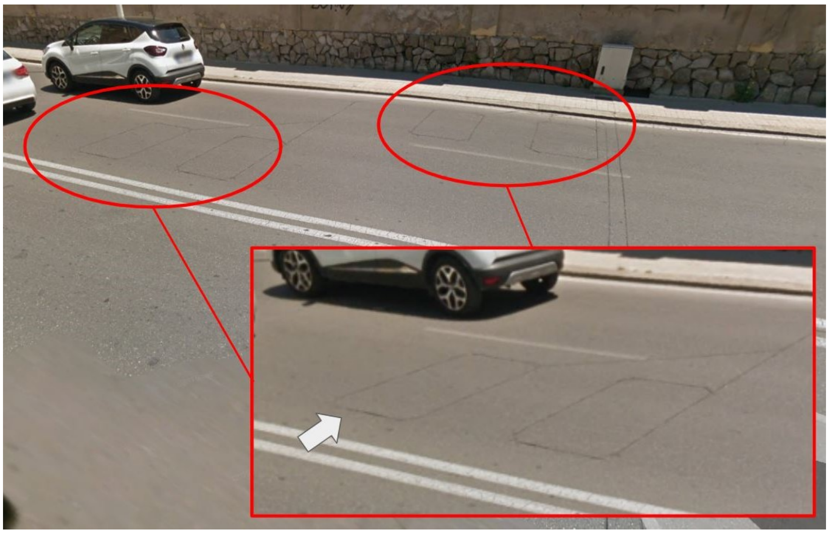


Figure 5.3: Inductive loops of traffic sensors installed in Cagliari.

of cars per hour that can be detected by that station. As the second step, the Prophet library[3] was used, which was created by Facebook for forecasting time series, to identify the remaining outliers and remove them from the data.

5.3.3 Data selection

At this point, the dataset has been cleaned up but has several missing data, arising not only from the removal of the anomalous values but also from errors in the system. (*i.e.* all those data that have never been saved on the database or have never even been sent due to a temporary sensor failure). It was therefore necessary to make a choice of stations to use before the next steps. All those stations that during the year 2020 have sent at least 60% of the data have therefore been selected.

5.3.4 Data Modelling

The last step in the data preparation pipeline is the data modelling, which is necessary for implementing a suitable data imputation strategy, studying seasonality and motifs present in the data. Again the Facebook Prophet library was used to build a clean and reliable dataset ready for analysis and visualisation (without any missing values or anomalies).

[3] https://facebook.github.io/prophet/

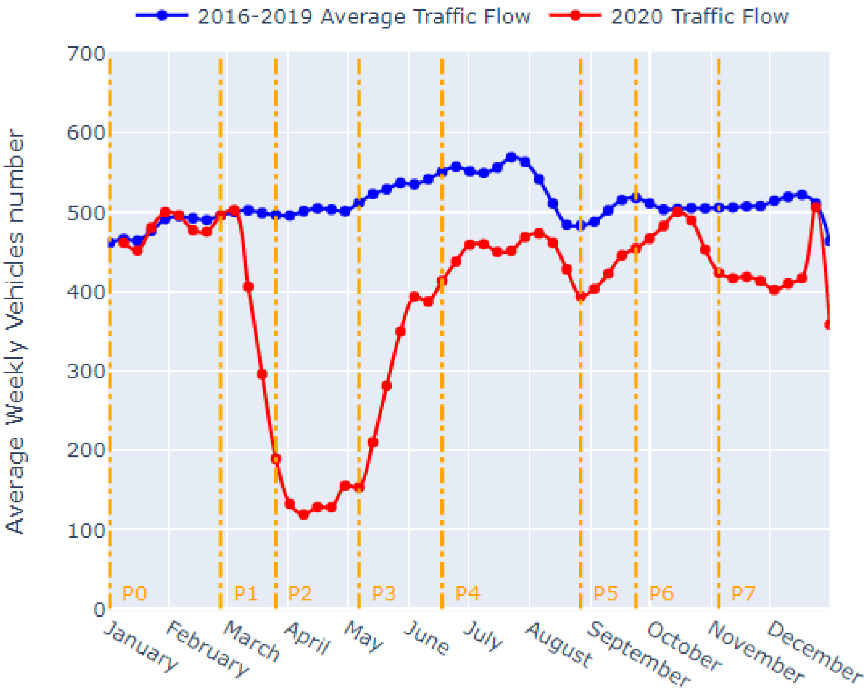


Figure 5.4: Comparison between 2020 and 2016-2019 average traffic volumes.

5.4 Data Analysis

5.4.1 City Traffic Analysis

In this section, an analysis of the traffic behaviour in terms of volume and pattern is shown. First, all traffic was aggregated to get a sense of the volumes involved and compare them against pre-pandemic levels. Figure 5.4 shows the traffic as an hourly vehicle average from January to December 2020, compared to average traffic flow from 2016 to 2019. Eight-time partitions (from $P0$ to $P7$) are used, which characterise significant patterns in traffic variability (*e.g.* significant changes or stable values). These are shown in Table 5.1, where the traffic variations are shown as a fraction of the pre-pandemic figures. The first lockdown is between $P0$ and $P1$ and the effect of subsequent mobility restrictions is also noticeable. After the first lockdown ($P3$) the traffic picked up rapidly, without reaching pre-pandemic levels during the Summer ($P4$), due to a significant decline in tourism. In period $P6$ there were again few restrictions and traffic went briefly back to pre-pandemic levels since October and November are non-touristic periods anyhow. Finally, the further fall in $P7$ reflects the second wave.

Table 5.1: Period details. The average is expressed as an hourly vehicles average.

	Average	Variation	From	To
P0	483.0	NaN	2020-01-02	2020-02-26
P1	277.0	-0.43	2020-02-27	2020-03-25
P2	145.0	-0.48	2020-03-26	2020-05-06
P3	368.0	1.54	2020-05-07	2020-06-17
P4	445.0	0.21	2020-06-18	2020-08-26
P5	444.0	-0.00	2020-08-27	2020-09-23
P6	463.0	0.04	2020-09-24	2020-11-04
P7	413.0	-0.11	2020-11-05	2020-12-31

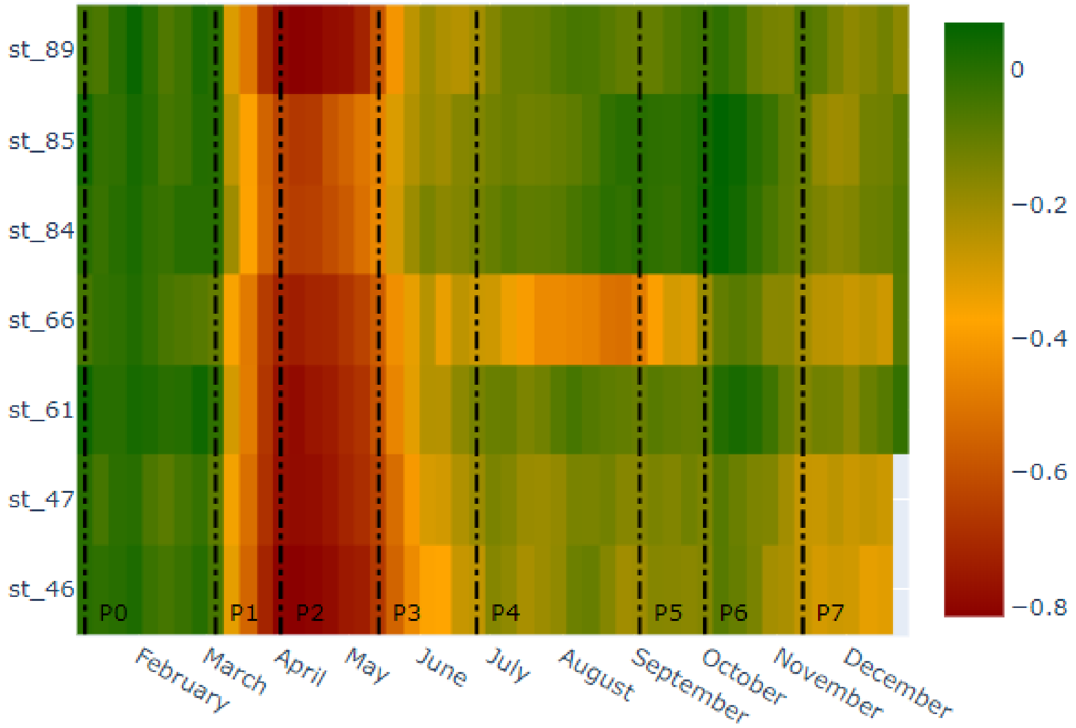


Figure 5.5: Variation of the 2020 traffic values as a fraction of the pre-pandemic volumes, averaged in 2016-2019.

In Figure 5.5 the traffic distribution patterns for each station under analysis have been analysed. These values express the relative changes between the pandemic-period traffic and the average of pre-pandemic years (2016-2019). This colourmap gives a different perspective, showing in green the period that correlates well with the past and then gradually turning yellow and red during the most striking changes. In fact, strong volume reductions are recorded in the $P2$ period (corresponding to the first lockdown) and then return to local maximums in the period between October and November. The most interesting thing about this graph is the different distribution across the sensors. Different stations have had different levels of traffic variations between them, indicating the traffic has shifted among different areas, in addition to getting absolute reductions.

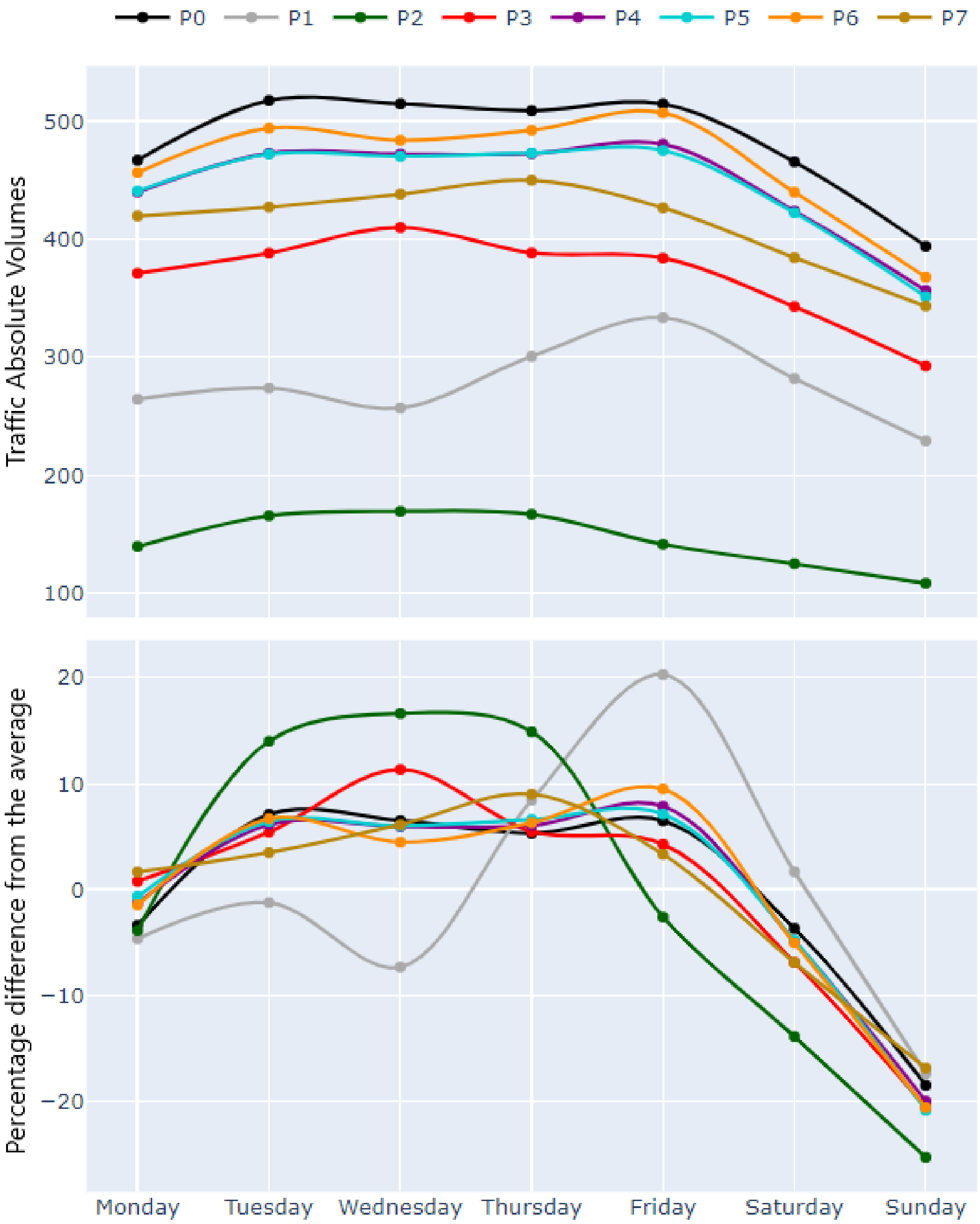


Figure 5.6: Weekly traffic distribution for each period, giving absolute values in the figure above and relative values in the figure below.

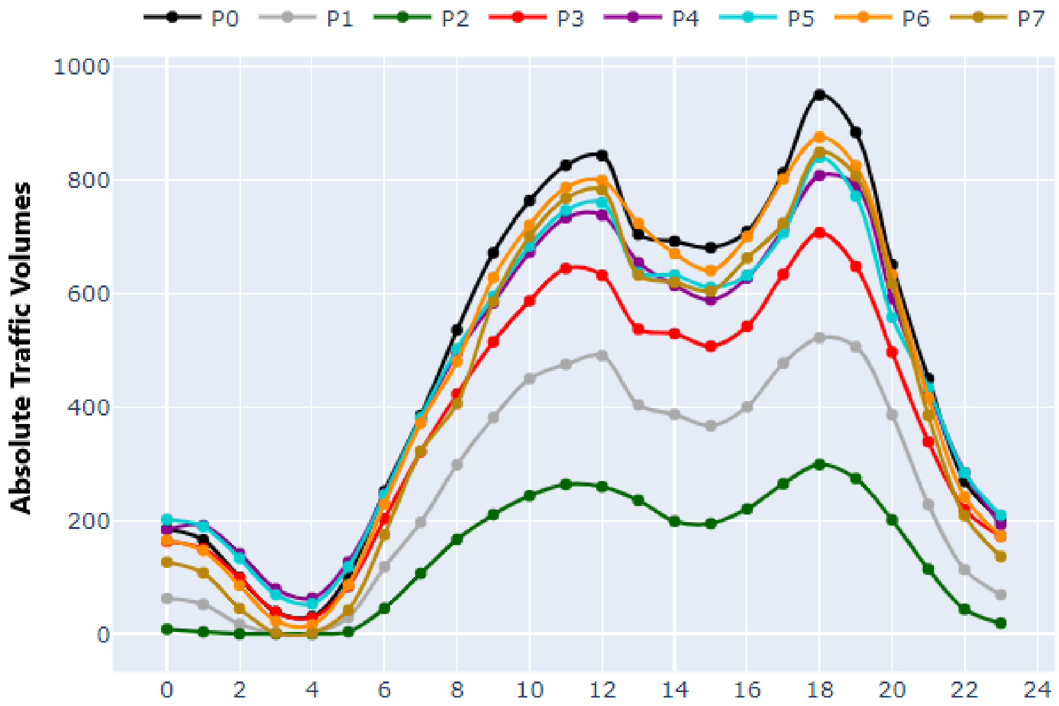


Figure 5.7: Daily traffic distribution for each of the eight pandemic periods.

On the other hand, Figure 5.6 shows the distribution of weekly traffic for each period, reporting the absolute values (above) and the relative change values (below). This shows striking changes in weekly mobility patterns during the eight phases of the 2020 pandemic year. In the pre-pandemic period of reference ($P0$), the traffic is fairly spread over all midweek days and then decreases over the weekend, which is typical of city traffic. During the start of the first full lockdown ($P1$) the weekly cycle is totally unprecedented, with a single peak on Friday. So there is not just a striking drop in traffic (-76%) but also a change in the habits. Then, once the lockdown is in full swing ($P2$), the traffic volumes remain steady; yet the weekly cycle gets closer to normal pre-COVID, but with a skewed anomaly on Fridays. The periods from $P4$ to $P7$ have a very similar pattern to the reference and also the volumes tend not to be too small, showing a readjustment to pre-COVID habits, albeit at lower volumes. On the other hand, the daily traffic cycles remain consistent during the whole pandemic year as shown in Figure 5.7.

5.4.2 Touristic Flows Analysis

Given the impact that the lockdown has had on general traffic, in this section it will be shown how it has affected tourism. The reference station chosen to carry out this analysis is station number 89, located on a 4-lane road that connects the city centre to the maritime part of the city where there is a long sandy beach, widely used in summer. The traffic flow in 2020 was reduced by 20.69%, and this value is in line with the official data recorded by the regional authorities. That means that by analysing the traffic flow in different places, it is possible to profile the behaviour of the citizens and the different traffic types. What shown so far is valid as general behaviour for the city. Analysing every single sensor or a cluster of sensors, it is possible to evaluate the behaviour of traffic in specific areas.

In Figure 5.8 the example shows the comparison between the distribution of traffic in a selection of sensors, in comparison with the curve of tourist presences in the city during the year. It can be seen that the distribution and trend are clearly all in line with the lockdown period. During the Summer (coinciding with low movement restrictions from July to September) two sensing stations differ strongly. Station 66 has a lower than average behaviour in those months. Yet this was found to be due to road-works. For station 89, however, the situation is different because the traffic flowing in this sensor correlates with the tourist presence curve, confirming this station as a strong indicator of touristic traffic.

5.5 Summary of key results

The research work presented in this chapter analyses the variations and impact in the mobility model that occurred in 2020 in the city of Cagliari, which is a particularly touristy coastal city. The entire analysis is based on the traffic data acquired by inductive loops sparse in the city. During the first lockdown period, traffic volumes dropped by up to 76% compared to average traffic values of the previous four years and then recovered in subsequent periods. Despite the presence of further restrictions at the end of the year in the vicinity of the second wave of infections, the traffic did not react proportionally.

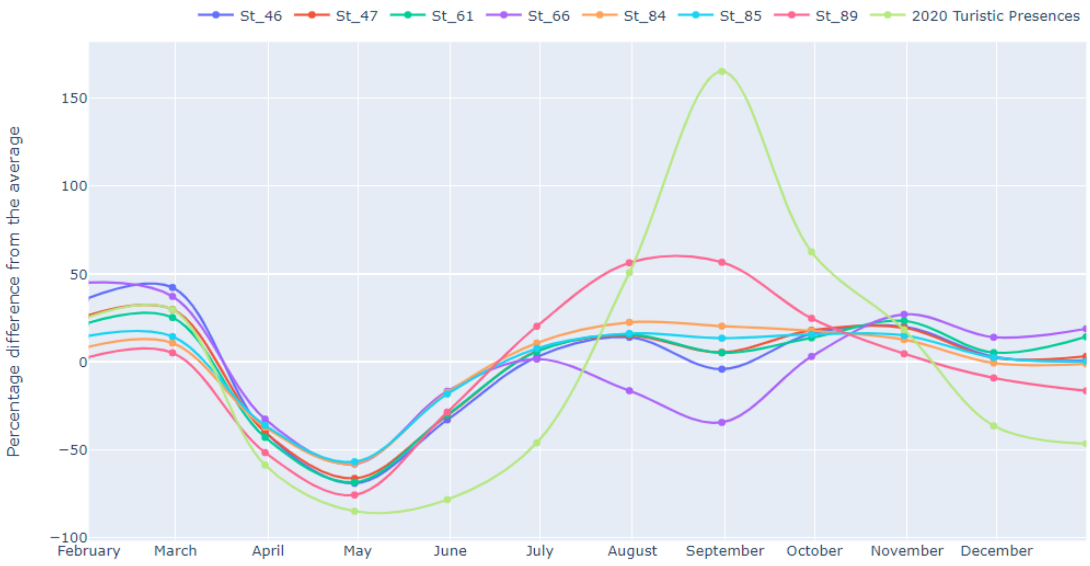


Figure 5.8: Comparison between 2020 traffic stations volumes and the touristic presences.

The change in traffic in the various stations was also analysed, where different behaviours emerged, relating to different areas. Through this pilot study it was possible to put real data to the test, unveiling a mix of expected results (volume drops) and less obvious results (change in traffic patterns). One could also evaluate the specific effect that the pandemic had on tourism, given that traffic volumes did not return to normal during the summer, in spite of more relaxed restrictions. This kind of analysis is of paramount importance in understanding the effect of mobility restrictions and may be further extended in many ways. It can be identified as an important contribution of this chapter, how it was possible to differentiate system anomalies (faults) from the actual mobility anomalies using two different steps of data cleaning.

Chapter 6

Discussion and Conclusion

This chapter concludes the entire thesis, reviewing the research presented. In Section 6.1 contributions made in Chapters 2, 3, 4, 5 and Appendix A were summarised. Finally, in Section 6.2 some further developments were discussed.

6.1 Summary of results

The primary aim of this thesis was to improve data collection and data analysis methods in urban scenarios by applying and developing data science techniques.

In Chapter 2 data science techniques and machine learning approaches to maximise insights regarding citizen's interaction with urban green areas, within a pilot study carried out in the Sheffield area (United Kingdom) were explored. The study data was collected via a smartphone app that simultaneously acquired subjective and objective data. After collection, the data underwent several stages of analysis. In the first phase, a deep cleaning process addressed the issues related to missing data, erroneous data or other elements that can affect the statistical significance of data. The second phase focused on the analysis of clean data, based on several metrics. Through the development of new analysis algorithms and also including machine learning techniques, it is possible first study the subjective data; *i.e.* the observations submitted by users.

The subjective and objective data was then merged; *i.e.* all the information extracted from the smartphone sensors (for example the position and movement speed) could be investigated. By applying these algorithms to the case study, the objective data was then analysed, which through the acquired location points made it possible to understand, for example, who are, not only, the most active users but also the interaction behaviours with parks and urban green areas. The work presented in Chapter 2 is therefore intended to inspire and support further social studies, as the lessons learned from this project led to a better understanding of how to conduct large-scale social studies and what techniques can be employed to obtain results from both objective and subjective data. Such studies may lead to optimised planning of services and targeted interventions in cities.

The work presented in Chapter 3 addressed some of the limitations of the approach used in Chapter 2. In active data collection (*i.e.* citizens become users and participate to generate subjective data of interest for the study), reducing intrusiveness and increasing the statistical significance of the data pose a significant challenge. In order to collect data on a large scale while improving the quantity and quality of the insights extracted, it is necessary to develop automatic frameworks. The search for statistical significance in urban sciences may require a departure from conventional methods, where data collection and data analysis phases are typically done separately. Instead, here data collection and data analysis were performed in parallel, using intelligent processes in real-time (during data collection) to guide the next steps of data collection.

Analysing user feedback in real-time (through AI-based feature extraction and text analysis) and combining feedback with context (position and level of information of each question in the test matrix subjective), led to a significant acceleration of the overall process. While a generic approach was developed (making the process applicable to different studies), in the real-world case presented, a 41% acceleration in achieving statistical significance and a reduction in intrusiveness by 75% was obtained. Comparable improvements are expected in similar studies involving both citizens and cities. These findings lay the foundation for the integration of intelligent data collection and analysis processes into urban studies.

In Chapter 4 the focus is on passive data collection. While results obtained using active data acquisition were promising and satisfactory, a paradigm shift in the data gathering approach was required, to limit the need for citizen involvement. This led to a passive data collection method which could extract data from a sample as large as possible. A low-cost device was therefore designed and developed to monitor and analyse people mobility in urban areas by exploiting the active scanning for access points discovery used in Wi-Fi protocol.

Through the acquisition of messages sent by smartphones, it is, therefore, possible to count and locate devices and consequently, with a low margin of error, people present within the monitoring area. Analysing the data collected with the acquisition sensor, it was possible to create different metrics, relating to crowd density, people flows moving between different areas, origin/destination matrix (useful for planning the transport service), stays and returns numbers. The solution was tested first under controlled conditions, and then in three pilot real-world deployments. In addition, two different approaches to locating people were tested, using algorithms based on RSSI and TDOA respectively.

In Chapters 2, 3 and 4 were focused mainly on data acquisition and analysis, the enormous amount of data already available through open sources required to deepen this approach too. The open data available can add significant value which should be considered and integrated into all the systems aiming to extract meaningful insights for services improvement in cities. In Chapter 5 the effective use of open data was investigated, based on data made available by the municipality of Cagliari (Italy), which reported the traffic data acquired by the inductive loops sensors scattered across the city. Based on this data, the variations in traffic patterns generated by the Covid-19 lockdown measures were analysed. This mobility study was particularly relevant for this city, which is located on the coast and is also a popular touristic destination. In this chapter the variations in traffic data between the average traffic values of the previous four years (taken as benchmark) and the traffic levels in 2020, subject to the restrictions imposed during various lockdown periods were analysed.

Through this pilot study involving real data, a mix of expected patterns (volume drops) and of peculiar trends (reassignment of traffic) was revealed, demonstrating the great potential offered by open data. This vast amount of readily available information can be exploited to significantly increase the value of the data collected and analysed in other ways (such as in the rest of this thesis).

Last but not least, in the Appendix a de-randomisation algorithm which allows to accurately discriminate which probe requests were sent by the same device, even when the messages were sent with a random MAC address was presented. Using this algorithm it is then possible to compute all metrics shown in Chapter 4, making it possible to collect data from devices strictly related to the number of people in the monitored area. This algorithm solves the delicate problem of detecting devices in a certain area, while respecting privacy constraints.

The contributions provided in this thesis advance the body of knowledge in people counting, localising and tracking. However, many interesting research developments remain to be pursued, detailed in the next section.

6.2 Future work

During this thesis work, several methods and frameworks were developed for the improvement of data acquisition and analysis in urban scenarios. For each of them the usability and limitations were discussed and the further steps required to overcome or address them were outlined.

In the active data collection approach presented in Chapter 2, it is possible, for example, to carry out new pilot projects to identify other characteristics that are critical for citizens wellbeing. Even more ambitious would be to explore the mutual interactions between citizens and “smart” cities, whereby more unrelated data must come together for the benefit of citizens. With the framework presented in Chapter 3 it is not yet possible to anticipate the necessary duration of a complex subjective study. It might be interesting to work on this and other pilot design factors to help plan and budget large-scale city pilot studies.

Regarding the passive collection approach, on can take advantage of the system presented in Chapter 4 and the related de-randomisation algorithm illustrated in appendix A. One can consider how fast the Wi-Fi protocol evolves and explore the new Wi-Fi 6, to both understand the compatibility and functionality of the proposed system, and to find any improvements given by the introduction of new parameters in the active scanning process of the new protocol.

Finally, regarding the mobility tracking presented in Chapter 5, one could aim to better differentiate the various types of traffic and to understand which types and volumes of traffic are essential for the normal functioning of the city: before, during and after exceptional events.

Pursuing these research avenues could further expand the literature on data science methods and techniques applied to urban scenarios.

Appendix A

A de-randomisation algorithm for Wi-Fi data collection

The material presented in this appendix is extracted from both the journal paper entitled "*MAC address de-randomization for WiFi device counting: combining temporal and content-based signatures*" submitted (and currently under review) at Computer Networks [118] from the conference paper entitled "*WiFi Probes sniffing: an Artificial Intelligence based approach for MAC addresses de-randomization*" sent to IEEE 25th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD) [13].

A.1 Introduction

To preserve people's privacy and prevent devices tracking, MAC addresses randomisation has been introduced by an ever-increasing number of operating systems. In Figure A.1 is shown a timeline that shows when the different OS manufacturers introduce that on their systems. As a result, mobile devices use *virtual* addresses that differ over time so that a single fixed *factory* address that can identify a specific user is not used. This has the consequence that it is not even possible to extract anonymous information on the mobility of people by analysing the traces of WiFi traffic, which would be useful for many purposes (e.g. counting the number of people in a means of mass transport) as illustrated in the chapter 4 where all metrics were based on counting and tracking only devices still sending their real MAC address.

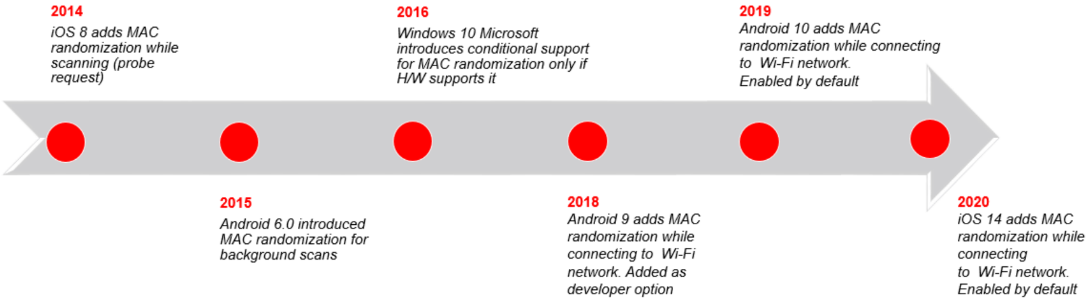

2014
iOS 8 adds MAC randomization while scanning (probe request)
2016
Windows 10 Microsoft introduces conditional support for MAC randomization only if H/W supports it
2019
Android 10 adds MAC randomization while connecting to Wi-Fi network. Enabled by default
2015
Android 6.0 introduced MAC randomization for background scans
2018
Android 9 adds MAC randomization while connecting to Wi-Fi network. Added as developer option
2020
iOS 14 adds MAC randomization while connecting to Wi-Fi network. Enabled by default

Figure A.1: Randomisation implementation timeline [119]

To address this issue, in this chapter a novel MAC address de-randomisation algorithm that groups specific messages (*i.e.* the Probe Requests) generated by the same physical device is proposed. With respect to the previous work, a features' combination that has been previously considered in isolation is considered, which are associated with the content and length of the optional fields conveyed in the sent frames and the rate at which the frame are numbered over time. These features are then used by density-based clustering algorithms (*i.e.* DBSCAN, OPTICS, HDBSCAN) to group frames sent by the same device.

Additionally, the presence of pseudo-random MAC addresses, which are those that do not change every frame but only when the emitting device switch on and off the WiFi interface was considered. To this, a heuristic to detect these sequences of frames so as to improve the algorithm efficacy was developed. Experiments have been initially performed in a controlled environment where an accuracy close to 96% was reached. Then, experiments in a real scenario have been conducted where the people taking the bus when moving in an urban area have been counted; in such a scenario an average accuracy of 75% has been obtained.

A.2 De-randomisation overview

The widespread use of connected personal devices (e.g., smartphones and smartwatches) have been exploited to develop appropriate strategies to monitor people's mobility by sniffing the generated data. One of the developed approaches relies on the use of the personal devices MAC addresses, which is captured from the device sent frames to track the mobility of the associated owners. In the last two decades, the ever-increasing privacy concerns have brought to the adoption of MAC address randomisation algorithms by an increasing number of operating systems. Accordingly, these devices make now often use different virtual addresses so that not a single fixed address is used that may identify a specific user. As a consequence, algorithms that intend to extract anonymous information (e.g., counting the number of devices in a given area) cannot similarly be effective anymore. Other consequences have been introduced and the IETF and the IEEE 802 standardisation committees are evaluating the impacts of these MAC address randomisation processes on existing use cases for network and application services. In fact, the MAC address was designed to be static and many services rely on this logic, but with randomisation, such services may not work properly anymore. This is the case of authentication procedures that may rely on the sender MAC address to keep the device authenticated when moving from one Access Point (AP) to another in an extended network.

Since 2014, when the first randomisation methods appeared on the market by smartphone and operating system manufacturers, several *de-randomisation* algorithms have been proposed, whose objective is to cluster frames generated by the same device and observed in a given period of time. Major proposals fall in the area of passive sniffing as they analyse frames generated by the devices when looking for APs they may use to connect to the Internet. These frames are then grouped and associated with different devices to be able to count the number of them in a given area (*i.e.* Probe Requests frames). The features that are considered are mostly related to the content of some optional fields that are conveyed in these frames, which vary from a device to another (e.g., [120, 121]; the temporal distribution of the sent frames (e.g., [122]); and the inter-frame time (e.g.,

[123]). The proposed algorithms are able to correctly group frames using different MAC addresses but belonging to the same device in up to 75% of the cases (best results). Whereas these results may be satisfactory for some applications domains, it has to be noted that the randomisation techniques introduced by the different vendors keep changing and are also becoming more and more complex. Accordingly, to keep these performance levels it is necessary to continuously test that the proposed features are still valid for the current traffic patterns and to look for additional features to keep the algorithm performance satisfactory.

Based on the previous considerations, in this chapter a novel passive sniffing de-randomisation algorithm is proposed. The provided contributions are the following:

- A combination of the features that have been used in isolation in past works and that are associated with the content and length of the optional fields conveyed in the sent frames and the rate at which the frames are numbered over time was considered. These features are then exploited by density-based clustering algorithms (*i.e.* DBSCAN, OPTICS, HDBSCAN) to group frames sent by the same device.
- The presence of pseudo-random MAC addresses, which are changed by the emitting device only it switches off and on the WiFi interface were considered. A heuristic to detect these sequences of frames (with almost static virtual MAC addresses) so as to improve the overall de-randomisation process efficacy was developed.
- the algorithm has been tested in a controlled environment; *i.e.* inside an anechoic chamber, so that the ground truth data was available. In this scenario, an accuracy close to 96% was reached, which is far higher than the performance achieved by previous works.
- an algorithm for counting the number of people inside a mass transport vehicles has been defined, which relies on filtering the frames on the basis of the received signal power. In such a scenario an average accuracy of 75% has

> been obtained. Whereas these results seem to be in line with previous works, it is worth mentioning that the increasing number and complexity of the randomisation algorithms adopted by the continuously evolving smartphone operating systems make these results a significant outcome. Additionally, this scenario suffers from the intrinsic error introduced by the fact that some people in the vehicle may either not have a device with a WiFi active interface or have more than one.

The rest of the appendix is organised as follows. Section A.3 provides the background information and briefly review the past works. Section A.5 describes the features that are used in the proposed de-randomisation algorithm, the clustering algorithm and the detection of the pseudo-random MAC. Section A.6presents the results obtained in a controlled environment. Section A.7 presents the algorithm that has been developed to count the people on board a mass transport vehicle. Section A.8provides final conclusions.

A.3 Background

The following subsections present background concepts and the state-of-the-art in the randomisation and de-randomisation procedures.

A.3.1 MAC structure, Probe Request frames and Information Element fields

Each WiFi device has a MAC address that uniquely identifies itself in the local network, whose structure is shown in Figure A.2. It consists of 48 binary digits, usually represented as a sequence of six octets of bits. The first 3 octets are assigned by the IEEE to producers and constitute the Organisation Unique Identifier (OUI). The remaining octets are called the Network Interface Controller (NIC) and are assigned by the manufacturer to each produced network card. This mechanism allows to identify the MAC address manufacturer and to assign each manufacturer with a unique address space from where take the needed addresses for each device. It is important to mention that the second last bit in the first octet, highlighted in Figure A.2, has an important role in addressing assignments. Specifically, if it is set to 0, then the MAC address is globally unique (according to the standard) and it is kept constant over time by the device; this is indeed the *real* address of the device. Otherwise, if set to 1, the MAC address should be locally administered with the consequence that the MAC address is randomly generated and may change over time. In this second case, the resulting address is considered as being *virtual* MAC. This behaviour is governed by the IEEE SA 802-2014[1] and 802c-2017[2] standards.

A device with a WiFi interface turned on sends special messages called *Probe Requests*, which have the objective to detect and identify the wireless networks that are available in the area where it is located. The Probe Requests are sent in groups called *burst*, whose length and inter-time vary from device to device. Each of them is associated with a time limit within which a specific message must be received to allow connection to the chosen network. The Probe Requests are received by the Access Points which in turn responds by sending a *Probe Response* frame containing the information needed to complete the association to the AP.

Both the Probe Request and the Probe Response frames convey some fields that are called *Information Elements* (IEs), which are of variable length. According to the relevant IEEE standard[3], the first octet of the Information Element encodes the Element ID, the second octet conveys the information element's length, and the remaining bits contain the relevant information in a number of bits equal to the length encoded in the previous octet. A variable number of IEs can be transmitted in a frame.

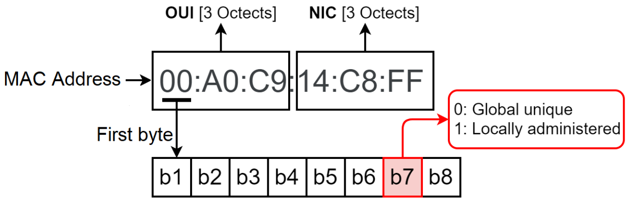


Figure A.2: MAC address structure. The seventh bit defines whether the address is of type "Global unique" or "Locally administered".

[1] https://standards.ieee.org/standard/802-2014.html

[2] https://standards.ieee.org/standard/802c-2017.html

[3] https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7786995

A.3.2 MAC randomisation

Table A.1: OS behaviour comparison for MAC address randomisation.

Randomisation behaviour across latest releases of Operating Systems			
Action	**Android Q**	**iOS 14**	**Windows 10**
Probe Mode Randomisation	ON. The users cannot change this setting.	ON. The users cannot change this setting.	OFF by default. The users can set it ON/OFF.
Randomise Daily	Optional	Not performed	Optional
Connection to an unknown SSID	On the first connection a random MAC is generated.	On the first connection a random MAC is generated.	*Randomisation ON*: On the first connection a random MAC is generated.*Randomisation OFF*: Factory MAC is used.
Connecting to a known SSID	When disconnecting and reconnecting, the same *virtual* MAC used for the first connection is used.	When disconnecting and reconnecting, the same *virtual* MAC used for the first connection is used.	*Randomisation ON*: When dis/re-connecting, the same *virtual* MAC used is used. *Randomisation OFF*: When dis/re-connecting, the factory MAC is used.
MAC Randomisation for a specific SSID Disabled	Device is automatically reconnected to SSID with factory MAC address.	Device is automatically reconnected to SSID with factory MAC address.	You need to manually reconnect to the SSID and the device uses the factory MAC address.
MAC Randomisation for all SSID Disabled	N/A	N/A	The device uses the factory MAC address.
SSID Profile Forget and Reconnection	If the SSID is forgotten, the device generates and uses a new random MAC address to connect.	If the SSID is forgotten, the device generates and uses a new random MAC address to connect.	*Randomisation ON*: the device generates and uses a new random MAC address to connect. *Randomisation OFF*: the device uses the factory MAC address to connect.

Currently, for privacy reasons, most of the devices in the active scanning phase; *i.e.* when the device sends the Probe Requests, hide their *factory* (also called *real* in the following) MAC address using a randomly generated one. This is done to guarantee that the real MAC address remains unknown so that the devices cannot be tracked over time. Accordingly, during the AP identification and connection establishment the device uses a virtual address and then switches to the real one once the connection is set up as starting from that moment the entire communication is encrypted. However, new policies have been introduced in the latest versions of some operating systems allowing devices not to reveal their real MAC address even when already connected to an AP. The MAC address randomisation process is regulated by IEEE standards; however, the algorithm implementation that generates them is different for each operating system.

In Table A.1 the behaviour of the Android Q, iOS 14 and Windows 10 operating systems with reference to the major action were compared. It can be noted that the randomisation process when sending Probe Requests is on in Android and iOS devices and cannot be turned off. These represent the OSs for the majority of smartphones making randomisation a common feature as the older versions of OSs in Android and MAC phones will soon disappear. Randomisation can be performed daily in some cases; *i.e.* the same address is kept for the whole day. As it appears from Table A.1, which MAC address is used depends also on whether the SSID is known to the device and whether the users have disabled the randomisation for some of all of these. Accordingly, it may happen that the same virtual MAC is used by the device when reconnecting to the same network. However, this new feature cannot be used to uniquely identify the devices under probing; on the contrary, it introduces an additional level of security to further safeguard the user's privacy.

A.3.3 Related past works

Many works have been proposed in the past to address the issue of MAC randomisation with several objectives, one of which is related to counting the number of devices in a given area. Most of these make use of a passive approach, which is the same exploited in this proposal. Some others make use of an active approach and finally, some perform a physical layer analysis. In the following, the most important works in these areas are reviewed and then the novelty of this proposal with respect to these past works are highlighted.

Passive sniffing methods

In [122], C. Matte et al. proposed a method to address the randomisation process by means of an algorithm that takes as input a set of Probe Request frames and associate them to the different devices. The principle is to create a signature based on the inter-frame time, the inter-burst time and on the frequency of frame sending; a measure of similarity between signatures based on the Franklin [120] distance is then defined. Finally, the MAC addresses whose similarity distance is below a certain threshold are aggregated. The resulting algorithm is recursive as at the end of each iteration it provides a list of groups with are again used to evaluate the distance of each frame with respect to the features of the different groups to eventually obtain a better association. A major weakness of this approach is that, due to some electromagnetic phenomena such as scattering or multi-path, the inter-frame times vary significantly between one burst of frames and the next, thus leading to multiple signatures for the same device. From the performed experiments, the resulting accuracy reaches 75%.

In the work presented by M. Nitti et al. in [123], a solution is proposed to count the number of passengers present in public transport vehicles. To identify whether two Probe Request frames have been issued by the same device, a score is computed which depends on the difference of the time of arrival and the difference of the sequence numbers. These differences are computed for all the possible couples of MAC addresses that have the same Information Element IDs, regardless of the length or content associated with it. Accordingly, the resulting algorithm assumes a recursive form and because of that is very computational intensive and it is very difficult to be used in a real scenario. The authors claim an accuracy of 100% in a controlled environment (closed room) and of 94% in a dynamic environment (a car). However, the tests have been performed with a limited number of devices and in a partially simulated environment. Also, [120] J. Franklin et al. proposed a timing-based approach; also in this case, the resulting solution is affected by errors due to uncontrollable physical phenomena, such as scattering and multipath.

Another work relying on Probe Requests fingerprinting is [121]. Herein, Vanhoef et al. have proposed a device tracking algorithm based on IE IDs fingerprinting. Their approach follows two phases. First, the IEs IDs are used to group Probe Requests into clusters, regardless of their temporal order and their content. Second, the algorithm tries to distinguish devices that are in the same cluster as they share the same IEs group. To do that, the algorithm relies on the predictable behaviour of the sequence number. To this, it assumes two probe requests belong to the same device if the difference in arrival times is lower than 500 seconds and their sequence number difference is less than 64. However, the probability of a device being successfully identified is less than 30% if the devices are more than 16. A corollary of this work is that pre-grouping probe requests based on their IEs is a good clue to find the MAC address pool which a device has changed over time and may reduce the computational cost of the previous approach, which was implemented in a recursive form, but paying a heavy price in terms of accuracy.

In [124], N. Suraweera et al. used WiFi packet sniffing to collect device-related compressed beamforming reports (CBRs). Indeed, downlink beamforming is facilitated by transmitting CBR from each wireless device to its AP. They exploited this information using the discrete 2-D Fourier transform (2D DFT) for feature extraction, similar to what is done for image matrices. The system was tested in a different environment from the training one and with devices that were not present during the training. The results obtained indicate 100% accuracy with no device, 97.8% with one device, 78.3% with two devices, and 93.9% with three devices present in the environment.

In [125], M. Ribeiro et al. give a counting and tracking method based on automatic classification techniques. This work was based on a 4-year probe request acquisition period. The collected data was inputted into 7 unsupervised classification algorithms; finally, the average accuracy was calculated by comparing the data collected by the authorities that administer the stadium, port and airport as ground truth. Thanks to the long observation period, the authors provided an overview of the rate of increase of devices randomising the MAC address. In particular, at the beginning of the acquisition, the devices that used the factory

MAC addresses were just over 50%, in the last period they are just under 5-10%. The authors applied the same method in [126] to create mobility tracks by monitoring passengers using public transport. Origin-destination matrices were created that provided an overview of which bus stops were most used by passengers. The analysis, in this case, was done without taking into account the random MAC addresses. Power filters were also applied to make it possible to understand which probes not to take into account for the analysis. The information extracted compared with the ground truth (number of tickets sold for the route) decreed that the proposed solution is not adequate to estimate the entrances and, consequently, the exits. However, it emerged that such information makes it possible to visualise and detect unusual situations and to raise awareness, support and facilitate communication between the different needs of stakeholders.

Active Sniffing Methods

The methods that belong to this category implement attacks that manipulate packets and fool devices on the network. In [121], Vanhoef et al. have analysed over 8 million probe requests with active techniques, counting around 170000 MAC addresses. The best-performing approach takes advantage of WiFi Protected Setup (WPS) parameters for devices that support it. The device being connected provides a parameter called Universally Unique Identifier-Enrollee (UUID-E). They found that this parameter is directly linked to the device's factory MAC address. This allowed the authors to trace the real MAC by greatly reducing the device count error.

Martin et al. [127] analyse various techniques that can be used on a large scale to be able to trace random MAC addresses to a single device. In particular, active sniffing methods exploit various vulnerabilities and made attacks such as KARMA attack [128] [129] (creation of fake Access Point from the list of probe BSSIDs) and RTS/CTS attack [130] in order to obtain the true MAC address during the negotiation of the connection with an AP, this approach requires the device's known SSID as knowledge of the attacker.

Physical Layer Analysis

The concept of fingerprinting has also been extended to level 1 of the ISO/OSI stack. In [131], V. Brik et al have employed machine learning tools to perform a fingerprinting of the device driver engaged in the transmission of IEEE 802.11 frames. This is possible by performing passive analysis of the radio frequencies used for data exchange. In laboratory experiments, they achieved 99% accuracy regarding device counting. However, in a real environment, such an approach is not very fruitful as the high interference would cause large inaccuracies and the data collection tools require complicated configurations.

A.4 The sniffer

Several improvements and changes were made from the sensor introduced in 4.3.1 PmA Stations, hence in this section the new sensor is presented again. The data acquisition focuses on capturing the Probe Request frames emitted by the WiFi clients in a given area. It is performed by a specially developed device called *sniffer*; it is shown in Figure A.3 and relies on the following hardware:

- 1 Raspberry Pi 3b+ with a custom firmware based on Raspbian Lite;
- 3 Wireless USB adapters with MT7601 chipset;
- 1 GNSS USB adapter;
- 1 LTE USB dongle to grant access to the Internet;
- 1 Li-ion battery.

A Python-based software module allows for collecting data over multiple WiFi channels through either fixed channel listening or channel hopping, as discussed in the following subsection. The software module architecture allows for storing the Probe Requests locally and sending them at regular and modifiable time intervals. In case there is no Internet connection available, the collected data is stored in an internal SQLite database, which is then sent when the connection is restored. Some pre-processing operations are also performed at the stations

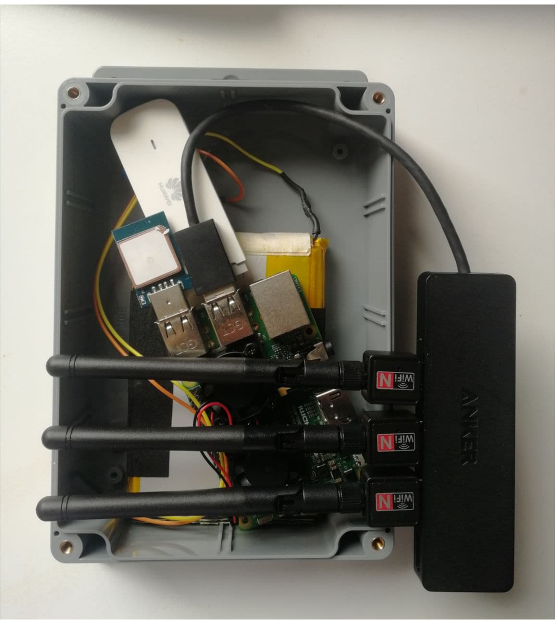


Figure A.3: Probe requests sensor with multiple interfaces.

to condition, compress, and clean the data. For instance, Probe Requests are eventually sent by APs and malformed packets are discarded (e.g., packets whose length declared differs from the actual length).

Another important operation implemented at the station is aimed at preserving users' privacy. Before sending the captured data to the cloud, the source MAC address is hashed with the PBKDF2 hashing algorithm which is one of the most resistant to "brute-force" attacks. We then store and use the resulting privacy-preserving hashes and discard the MAC address, which constitutes personal data. WiFi clients may follow two ways to access the network: passive scanning and active probing [132]. According to the passive mode, APs broadcast beacons packets to signal their presence, and the clients listen on channels for a fixed period of time. This approach, although completely passive, has a negative impact on the AP discovery process duration because the client needs more than 1.1 seconds to listen in all channels if an AP is present [133]. Instead, in the active probing mode, the clients continuously send Probe Request packets to discover the presence of APs and when doing it they hop over the used channels with a brief pause between them (a typical pause value is between 10 and 50 ms) [133]. As a consequence, the active mode is shorter and for this reason, it is the preferred

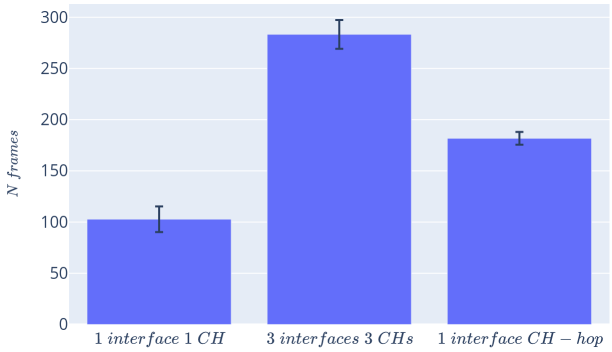


Figure A.4: Number of frames captured using three different sniffing configurations: a single interface with a fixed channel, three interfaces in three fixed channels, and a single interface with channel hopping (frame frames generated by a single smartphone in a semi-anechoic chamber).

approach.

An important mechanism that must be investigated when implementing a WiFi sniffer is the use of multiple monitoring interfaces and the eventual implementation of channel hopping. Like many other, not off-the-shelf WiFi sniffers, also the first version of the proposed sniffer had only one interface. However, it is, in general, true that increasing the number of antennas leads to an increase in the total number of packets captured. The ideal situation is when 14 interfaces listen on the different channels in parallel; *i.e.* one interface per channel for 100% of the time.

To evaluate the benefits and drawbacks of different configurations, extensive experiments were conducted to analyse the number of packets captured using 1 interface or 3 interfaces for the sniffing operations. Herein, the results of a specific test that was carried out analysing the total number of frames sent by a single smartphone (Honor9 with Android9.0) for 5 hours inside a semi-anechoic chamber. To acquire all the data, a sniffer with 4 antennas was used. One antenna was set to acquire data on all channels by performing either fixed sniffing in channel 1 (*1Interface1CH*) or channel hopping with 1 second of permanence time per channel (*1InterfaceCH-hop*), whereas the other 3 were set to acquire packets on fixed channels, specifically channels 1, 6 and 11 (*3Interface3CHs*). Figure A.4 shows the results in terms of the total number of Probe Request frames acquired by the station with the three configurations. It can be noted that an increase in the number of used interfaces leads to a higher number of captured frames; specifically, around three times the number of frames have been captured with the three interfaces sniffing in parallel with respect to the case with only one interface. When using the channel hopping technique an intermediate result is achieved and it is this approach that is used in the developed sniffer.

A.5 The proposed de-randomisation algorithm

The proposed algorithm is applied to a trace of Probe Request frames that are captured during an observation window of length T^W. As it will be shown in the performance analysis section, the longer the observation window the better the performance of the proposed algorithm to correctly aggregate captured frames generated from the same device. On the other hand, the shorter the observation window the higher the temporal resolution of the monitored varying number of devices located in the area of interest. The number of captured frames heavily depends on the number of devices, the artefacts present in the area and the gain of the antenna; it ranges from 50 to 200 when $T^W = 10\,\text{min}$ considering the whole spectrum. The first operation performed is data cleaning, which is aimed at checking for data errors, which may be introduced by the acquisition phase; deleting unnecessary information which would unnecessarily increase the computation burden; and introducing some corrections in case of missing values.

The data cleaning process starts by discarding all the corrupted packets to avoid the introduction of errors or false information in the following operations. On the resulting trace, a first filtering is applied that is aimed at isolating the probes sent by devices that use their real MAC addresses. It is performed by checking the $7th$ less significant bit of the $1st$ octet of the MAC address, as it has been shown in Figure A.2. As a result, two subsets are obtained: the first one of packets with real MAC addresses, from which it is straightforward to count the relevant devices; and the other one is further processed to extract meaningful information. To further process this trace, it is possible to rely on some empirical observations that emerged from the analysis of several datasets:

- Most devices that implement the randomisation process generate burst probe frame sequences where the virtual address is kept constant. In addition, the same device keeps IE IDs and the associated information content lengths constant over different bursts, even if the virtual MAC address changes. An example of this phenomenon is shown in Figure A.5, where frames of only two devices are considered for an observation window of 60 min. In the graph, a burst of frames is represented by a green or red bullet. The frames within each burst have the same virtual MAC in the source address. A burst detail is also shown in order to visualise the individual frames and their SEQs and TOAs variation. It has been possible to group the bursts in the two categories red and green associated with two different physical devices by checking the LENs of each IE ID for each frame. This has been possible because the two devices do not share the ID fingerprint. On the bottom of the figure, the used IE IDs are shown for the two devices and the associated virtual MAC addresses that have been used.

- However, when many devices are in the same acquisition area, it happens that they share the same fingerprint and for this reason, it is not enough to use only the ID fingerprint to discriminate the devices but a more complete analysis is required which considers the *speed* at which the sequence number field in the frames is incremented over time. Each device uses a speed that often characterises its Request Probes generation process. This feature needs to be used together with the ID fingerprint.

- Some other devices do not generate a new virtual MAC address every burst but only when they switch off and on their WiFi interface. These devices are easier to be detected and counted. If these sequences are detected, it is possible to make easier the identification of the devices for the remaining of the acquired trace. However, it is not always straightforward to identify the beginning and last frames sent by the same source. In the following, these devices are called *pseudo-random* devices.

These considerations have been used to develop the proposed algorithm, whose

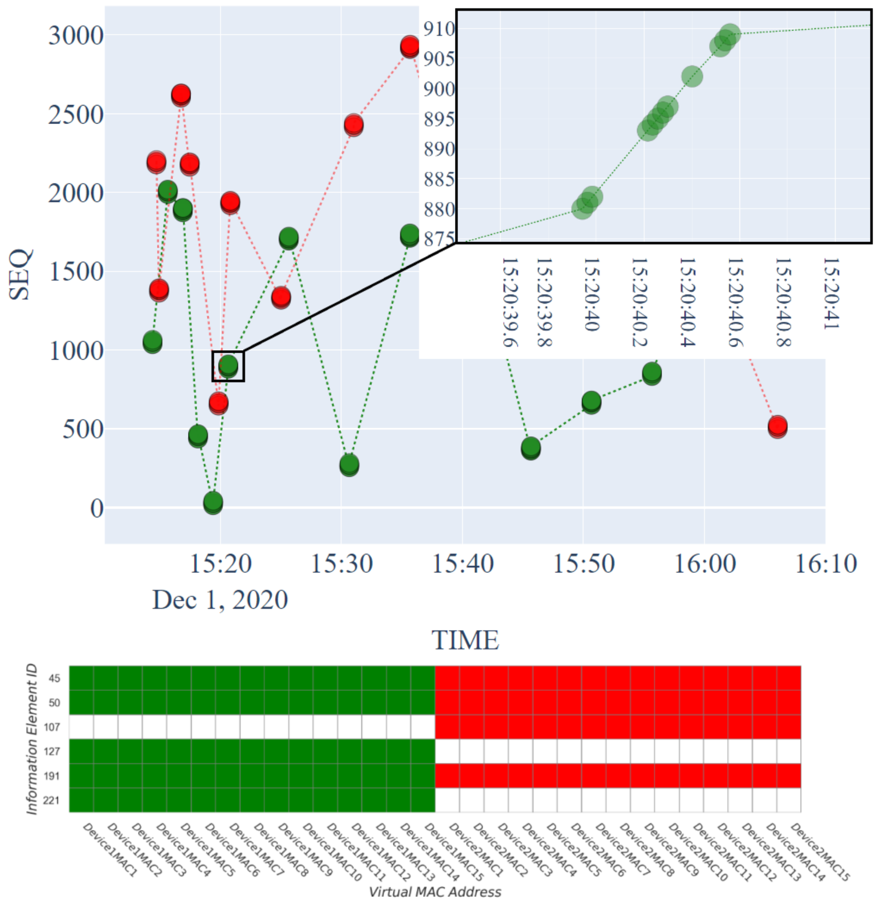


Figure A.5: Sequence of frames generated by two devices in an observation window of 60 min with SEQs and IE IDs details.

workflow is provided in Figure A.6. In the following subsections the feature extraction process, the filtering of the pseudo-random frames, and the clustering algorithm are described. To make easier the reading, in Table A.2 the notation used is summarised.

Table A.2: Used notation.

Parameter	Meaning
T^W	Observation window
n	Index of the number of unique MAC address during T^W
M	Sparse matrix with probe request details
m	Number of IE IDs found in all captured packets
InF	Time between two consecutive frames
TOA	Arrival time for a frame
SEQ	Frame sequence number
k	Specific burst
i	Number of frame in a burst k
t, g	Used for point P to compute the burst rate
j	Specific virtual MAC address
l	Length of burst k with MAC address j
Θ	Cumulative time of all burst
Λ	Number of burst with MAC address j
Ψ	Percentage burst time presence
χ_i	Internal parameter for Optics clustering algorithm

A.5.1 Features extraction

As mentioned before, the features that have been found to be relevant for the proposed algorithm are linked to the IE ID, IE length, the sequence number and the time at which the probes are generated. In the following, the trace processing procedures to explain the final features used is described. The first step consists in removing the unnecessary information from the captured traces and representing the remaining data in the most convenient way. Specifically, the fields that are kept from each frame are IE IDs and the IE length (called LEN in the following). Note that the IE content is not kept as this has demonstrated not to convey any additional distinctive information for the study purposes with respect to the LEN field. Together with these fields, the MAC address, the probe request's Time of Arrival (TOA), the sequence number (SEQ) and the received signal strength indicator (RSSI) are stored.

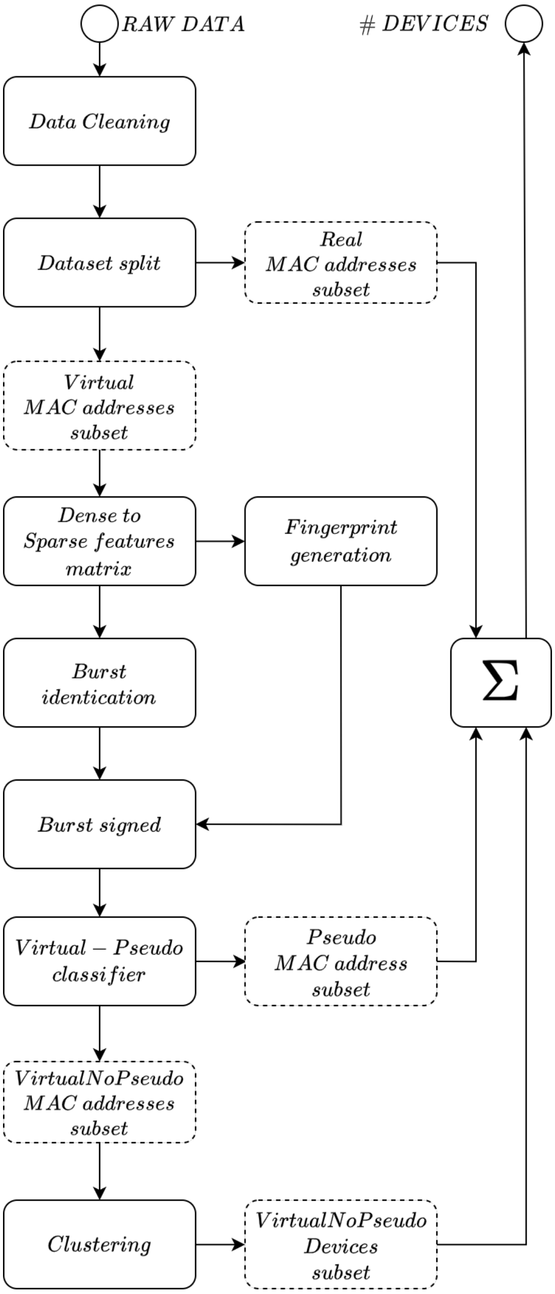

RAW DATA
# DEVICES
Data Cleaning
Dataset split
Real MAC addresses subset
Virtual MAC addresses subset
Dense to Sparse features matrix
Fingerprint generation
Burst identication
Σ
Burst signed
Virtual – Pseudo classifier
Pseudo MAC address subset
VirtualNoPseudo MAC addresses subset
Clustering
VirtualNoPseudo Devices subset

Figure A.6: De-randomisation algorithm workflow.

An example of this information is shown in Table A.3. For each frame captured, more than one row can be present as it is the case in the considered example for frame with $SEQ = 250$. For convenience in processing by keeping the same information, this table is converted into a sparse matrix as shown in Table A.4, which highlights that there is a column for each IE ID associated with the corresponding length. As some Probe Requests do not convey some IEs, the corresponding content is not available ("NaN" value in the examples of Table A.4). The resulting matrix has the following size.

Table A.3: Example of information extracted from the Probe Requests.

MAC	**Time of Arrival**	**SEQ**	**IE ID**	**LEN**	**RSSI**
MAC_1	1604571222.4325	250	45	5	-63
MAC_1	1604571222.4325	250	50	8	-63
MAC_1	1604571222.4325	250	221	6	-63
MAC_2	1604571281.9861	3500	127	10	-41

$$\mathbf{M} \in \mathbb{R}^{n \times m} \tag{A.1}$$

where n is the total number of different Probe Request frames received in the observation window T^W and m is the total number of different IE IDs found overall the received messages plus 4 (due to the columns for the fields RSSI, SEQ, MAC and TOA). Due to the variegate set of devices that have emitted the captured frames, each row has different IE columns with no information. By analysing different datasets, it was noticed that the frames sent by the same device contain a constant list of IEs and with constant lengths, as already discussed.

Table A.4: Features extracted from Probe Request packets as sparse matrix.

MAC	**TOA**	**SEQ**	**RSSI**	**ID 45**	**ID 50**	**ID 127**	**ID 221**
MAC_1	...	250	-63	5	8	NaN	6
MAC_2	...	120	-41	NaN	NaN	10	NaN

Furthermore, the different IE IDs have different importance for the objective of these analysis as some are more frequently used. For instance, Table A.5 shows the IE IDs that have been observed in a typical two-hour-long capture and that are used by the devices at least 5% of the total number of frames. On the basis

Table A.5: Presence rate of the IEs on the frame total number.

IE ID	Frames number	Presence Rate
50	24663	0.998
45	14953	0.605
3	13779	0.557
127	11979	0.485
0	10519	0.426
221	8659	0.350
107	2031	0.082
191	1635	0.066

of this feature, it appears that the above-mentioned IDs it is possible to identify a unique fingerprint for each device that uses a different virtual address over separate bursts. This aspect was also analysed in previous works, e.g. [134, 13]. Accordingly, a smaller matrix with only the mentioned 8 IE IDs is taken into account:

$$\hat{\mathbf{M}} \in \mathbb{R}^{n\times 12} \tag{A.2}$$

In this way, considering only 8 IE IDs, it is possible to reduce the features space and obtain an improvement in the performance of the clustering algorithm. The following operations are then performed:

- *Burst identification:* the objective is to identify all the bursts and to give an identifier. A burst is defined as a sequence of frames with the same MAC address with inter-frame time always shorter than a given threshold, called maximum interframe time t^{InF} (a typical value for it is 0.5 sec). A subsequent burst begins as soon as there is an inter-frame time that is greater than t^{InF}.

- *Fingerprint generation:* in this step the fingerprints based on the IEs are created. To this, starting from $\hat{\mathbf{M}}$, the whole set of IDs and LENs combinations associated with the different MAC addresses are identified and labelled with a unique identifier (named *fingerprintID*). Additionally, when some MAC addresses were observed to use multiple fingerprints (rarely) there were discarded not to create ambiguity in the analysis. In general, multiple fingerprints are interpreted by the algorithm as multiple devices.

- *Computation of burst rates:* at this stage, all the frames are grouped by burst and an aggregation operation is performed to extract the descriptive characteristics of each burst. Then the *burst rate* is computed, as a representative characteristic of the single burst. To this let define $t_{k,i}^{TOA}$ and $g_{k,i}^{SEQ}$ as the TOA and SEQ number of frame i in burst k, respectively. Points $\mathbf{P_{k,i}} = (t_{k,i}^{TOA}, g_{k,i}^{SEQ})$ for burst k are then represented in a plane and a linear regression is computed. The angular coefficient of the resulting line is the burst rate, which is a distinctive feature of each burst and represented by P_k. This feature has been empirically observed to be similar for bursts generated by the same devices and it is used for burst clustering.

Figure A.7 shows an exemplary capture lasting 25 min where each dot represents a distinctive burst of frames with the same MAC address. Different colour has been assigned to each different MAC address. It is possible detect visually that there is at least one pseudo-random device (violet dots) whereas some random devices can be easily detected as the sequence numbers increase regularly drawing a line on the 2D plane.

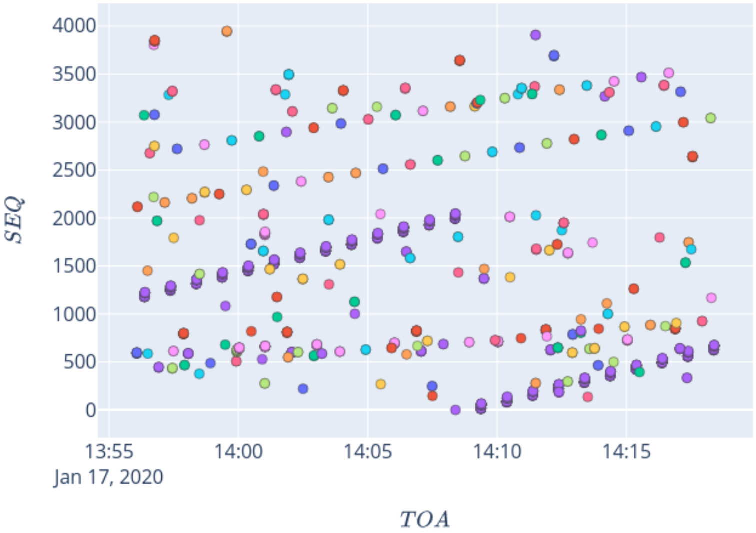


Figure A.7: 25 minutes capture of Probe Requests: each dot represents a burst of frames, different colours means different MAC addresses.

A.5.2 Pseudo-random MAC filtering

As mentioned before, pseudo-random devices are those that use the same virtual address for many bursts. It is then appropriate to identify these devices and to remove the generated bursts before applying the clustering operation. The reason is twofold: i) the fact that the same MAC is kept constant in a group of bursts makes this operation simple and reduces the complexity of the next clustering operation; ii) they can create some noise in the clustering and reduce its performance. However, the identification of these bursts is not that straightforward as there is not a fixed length of this sequence of bursts and the MAC address used by the pseudo-virtual device may be also used by other virtual address devices. Accordingly, a specific procedure had to be devised. Three metrics were considered:

- the *cumulative time of all bursts* having the same virtual address j

  $$\Theta_j = \sum_k l_{k,j} \tag{A.3}$$

  where $l_{k,j}$ is the *length of burst* k having the same MAC address j;

- the number of bursts Λ_j with the same MAC address j;

- the *percentage of time in the observation window* T^W during which the bursts are observed which

  $$\Psi_j = \frac{t^{TOA}_{k,last} - t^{TOA}_{k,first}}{T^W} \tag{A.4}$$

  where $t^{TOA}_{k,first}$ and $t^{TOA}_{k,last}$ are the TOA of first and last frames having the same MAC address in T^W.

The higher the value of these parameters the higher the probability that these bursts have been generated by a pseudo-random device. These features have been used as features space for a Support Vector Machine (SVM) binary classifier [135], which has been trained using long traces where the pseudo-random devices were known. Figure A.8 shows the Receiver Operating Characteristic (ROC) curve

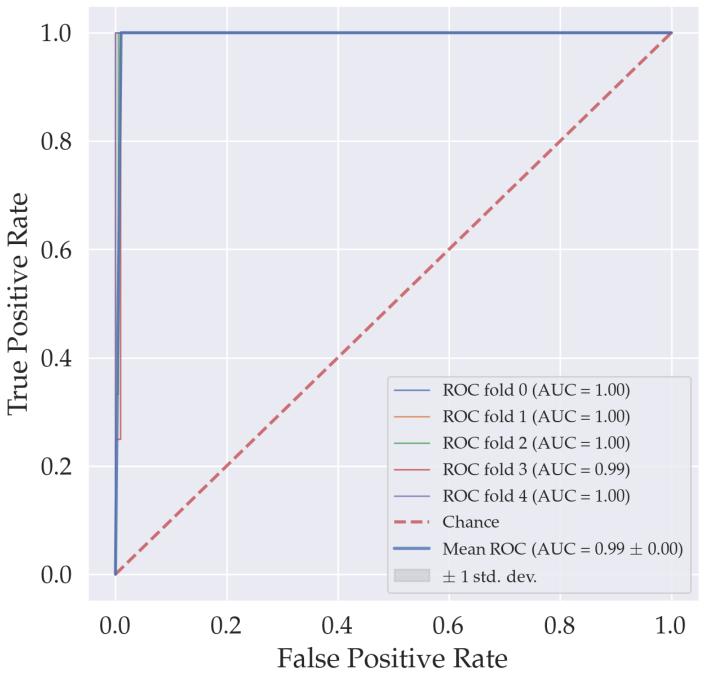


Figure A.8: Receiver Operating Characteristic (ROC) curve of the SVM classifier applied to the identification of the Pseudo Virtual frames.

As it is well known, when the curve reaches the top left corner of the plot very good results are achieved; *i.e.* a false positive rate of zero and a true positive rate of one. The chance line represents the ROC curve when the classifier has an equal probability to predict correctly or wrongly. From the mean ROC curve, it is possible to deduce how the chosen features are able to divide the samples well into the two classes of belonging (Virtual and Pseudo Virtual), reaching 100% accuracy values in almost all the 5 iterations of the k-fold cross-validation.

A.5.3 Frame clustering

The output of the previous operations is the division of the trace into frames with *real*, *pseudo-random* and *virtual* source addresses. For the first two groups, counting the number of devices is straightforward. For the third group, a proposer clustering operation has to be implemented. To do this, the following features are used for the identified bursts: the *burst rate*, the set of IE's LENs and the average inter-frame time InF_k of the burst k. Using the characteristics obtained, it is possible to proceed with the clustering of the bursts. Among the possible approaches, the one based on density has been selected as it has the advantage of being able to create arbitrary clusters and it can reach a high level of scalability if properly configured.

Among the density-based algorithms, there is a well-known algorithm which is called Density Based Spatial Clustering of Application with Noise (DBSCAN) [136] and its Hierarchical version HDBSCAN [137]. Other alternative solutions have also been proposed, such as the Fuzzy Joint Points (FJP) [138] and the Noise-Robust Fuzzy Joint Points (NRFJP) algorithms; however, these methods suffer from the low speed of the FJP algorithm. Therefore, it is not recommended for use in those applications which need to manage large datasets. Finally, the a successor of the DBSCAN has been proposed; *i.e.* the OPTICS algorithm (Ordering Points To Identify the Clustering Structure) [139].

Density-based clusters are defined in the features space as variable density areas separated each other by more rarefied areas. The idea could be explained introducing the definition of *core-points*, *density-reachable points*, *density-connected* and *outliers* or *noise*. A point is a core point if its neighbourhood of radius ε contains at least $MinPts$ points. A point u is defined a *directly-density-reachable point* if it is in the neighbourhood of a core point p. However, a point could be only *density-reachable* if there is a transitive closure of direct density-reachability; *i.e.* if there is a third point r from which both u and p are density-reachable. Finally, outliers are defined as the set of points in the dataset which do not belong to any cluster.

In this work DBSCAN, HDBSCAN, OPTICS algorithms were adopted, which have the following specific features:

- $DBSCAN$: once ε and $MinPts$ are given, the clusters' density is defined and is not possible to change it during the clustering process. In the experiments done for this thesis, good results were found with $MinPts = 2$ and ε varying from 0.0001 to 0.1.

- $OPTICS$: it is based on the principles of DBSCAN and follows all its definitions, but addresses one of the major weaknesses of DBSCAN; *i.e.* finding important clusters in the data varying the density threshold ε. To this, the samples are linearly ordered in such a way that spatially close points become neighbours. Putting these points in a x-y plane with the ordered points in the x-axis and the reachability-distance on the y-axis, it is possible to see a *reachability plot*. Such kind of plot allows to see different density clusters and to calibrate the boundary of a cluster simply by using the value of its derivative χ_i as a threshold. In this case its value is has been set to 0.1.

- *HDBSCAN*: Hierarchical DBSCN has been developed by Campello, Moulavi, and Sander [137], with the objective to provide only a *flat* (*i.e.* non-hierarchical) labelling of the patterns, based on the global density threshold ε. This allows HDBSCAN to find clusters with different densities (unlike DBSCAN), and be more robust to parameter selection.

Our features space is composed of all the LENs, the median inter-frame time and the rate burst. Therefore, by applying the clustering algorithms it is possible to discriminate the different devices. The comparison of the results with the various clustering algorithms is shown in section A.7.3, in fact, once the clustering has been finally performed, the algorithm has produced three subsets of devices:

- Devices with a Real MAC address.
- Devices with a Pseudo Virtual MAC address.
- Devices with a set of Virtual MAC addresses.

The sum of elements belonging to each subset gives us the estimation of the total number of unique devices which have been observed in T^W in the area where the capture has been performed. However, the lists are kept separate as other metrics can be derived, like those that rely on the identification of devices that come back after an interval of time for the devices which has a real MAC address.

A.6 Results in a controlled environment

To analyse the performance of the devised de-randomisation algorithm, the experiments were conducted isolating the test devices from other external ones. For this reason, the data acquisition sessions were conducted in a semi-anechoic chamber located at the Faculty of Engineering of the University of Cagliari. Figure A.10 shows the setup with all the test devices and the laptop that implemented the sniffing process and the proposed algorithm for de-randomisation. In this experiment, only one interface was used which was sniffing on channel 1. Table A.6 lists the number of test devices with the operating systems.

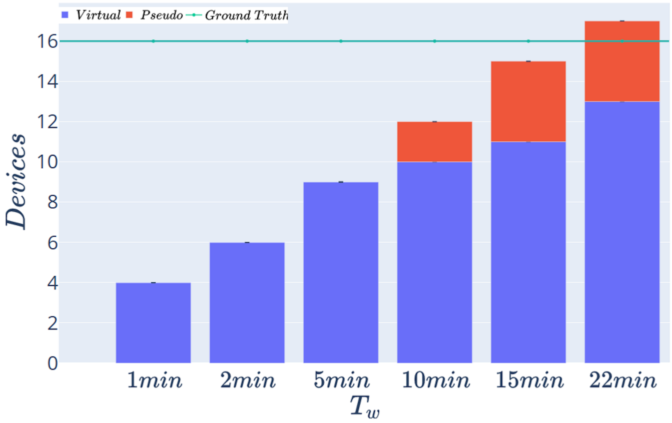


Figure A.9: Semi-anechoic chamber tests: number of devices counted with the proposed algorithm varying the observation window.

Figure A.9 shows the results at varying observation window T^W. As expected, the longer T^W the better the performance. The reason is twofold: when the observation time is short, the station is not able to capture any frame from some of the test devices; when only few frames are captures from each station, it is not possible to compute the *burst rate* feature that is used to separate frames emitted by different stations and that are then separated by means of this feature.

Indeed, note that at short observation windows the algorithm under-counts the number of devices. It can also be observed that an observation window T^W of at least 10 minutes is needed to start distinguishing devices with pseudo-virtual MAC addresses from those with a virtual MAC address. The main reason is that in 10 minutes the algorithm is not able to identify repetitions of runs with the same MAC address. The best result obtained in this scenario is with T^W set to 22 minutes, reaching an accuracy of 97% using DBSCAN as core clustering algorithm, which means that the error is less than 0.5 devices over a total of 16 test devices. This result represents an excellent basis to build solutions that can work in real scenarios. For this reason, further development and testing of a solution for an Automatic Passenger Counting system to be used on-board of buses has been carried out. Note that in this test the devices using real MAC addressed have not been included in the analysis but these have been excluded as soon as these have been detected.

Table A.6: Test devices used in the semi-anechoic chamber test.

BRAND	**# of devices**	**OSs**
Huawei	5	Android 5.1(1)/6(1)/8(2)/9(1)
Samsung	6	Android 4.2(1)/7(1)/8(1)/9(3)
ZTE	1	Android 8
Xiaomi	4	Android 6.1(1)/9(3)
Honor	1	Android 9.1
Apple	5	iOS 12.4.4(1)/13.3(4)
Motorola	1	Android 7

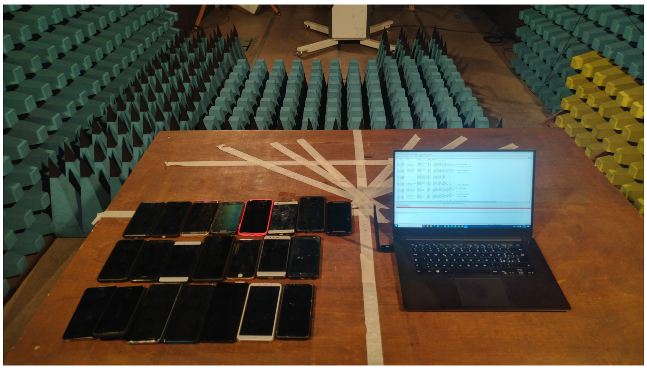


Figure A.10: Setup of the anechoic chamber tests.

A.7 Results on-board of city buses

This section illustrates the results obtained when using the proposed algorithm to develop an Automatic Passenger Counting system. The following subsections present the data acquisition setup, the analysis of the data and the counting performance.

Table A.7: Details of the Datasets used.

Dataset **Name**	Anechoic Chamber	Line 1 Brotzu	Line 1 Gioia	Line 30 Brigata Sassari	Line 30 Matteotti	**Totals**
Length [mins]	32	22	48	28	34	164
#Pkts	9707	15589	27107	22568	25693	100664
#MACs	300	1187	2032	1101	1695	6315
#Real MACs	31	363	524	285	278	1481
#Real Pkts	3598	6285	12235	12610	11955	46683
#Virtual MACs	269	823	1507	815	1416	4830
#Virtual Pkts	6109	9184	14504	9887	13579	53263
Ground Truth	16(v) + 8(r)	32	45	20	43	

A.7.1 Data acquisition setup

We have installed three sniffers on-board of three different buses of the CTM local public transport service company that operated in the city of Cagliari, Italy. The buses were scheduled to work on different service lines every day. This allowed us to acquire a significant amount of data on the behaviour of travellers when using urban transport services in the different areas of the city. In each bus under test, a sniffer was installed above the central door to better cover the whole bus area.

As for the power supply, different configurations have been tested. In the first trial, the sniffer's power supply was connected to the onboard services power line; thus, when the bus engine turns off, the power line is also switched off causing the sniffer to terminate the acquisition abruptly. This introduced a problem when the bus reaches the last stop where the engine is turned off while people continue to get on and off the bus. On the other hand, using the power lines under the battery is not a solution because the sniffer could drain the battery if the vehicle is not put into service for a long time.

By studying the electrical system of the bus, a viable solution was identified which is presented by connecting the sniffer power line to the power line of the ticket machine; indeed, this line is stabilised and it has a shutdown time which is delayed of 20 minutes after the bus engine is turned off. Accordingly, the sniffer could acquire the data of interest even at the last bus stop, when the bus engine is off, and at the same time, it turns off after 20 minutes once the bus has been turned off at the end of the service, preventing the battery from being discharged.

The acquisition took place from July 2020 to October 2020. Among the several data acquisitions collected through the sniffers installed onboard the buses, two specific bus service lines were chosen to be analysed with tests performed in both directions. Specifically, line 1 (directions Brotzu and Gioia), which is one of the longest lines, and line 30 (directions BrigataSassari and Matteotti), which is the service line that connects the city with the second-most populous city in the metropolitan area, were selected.

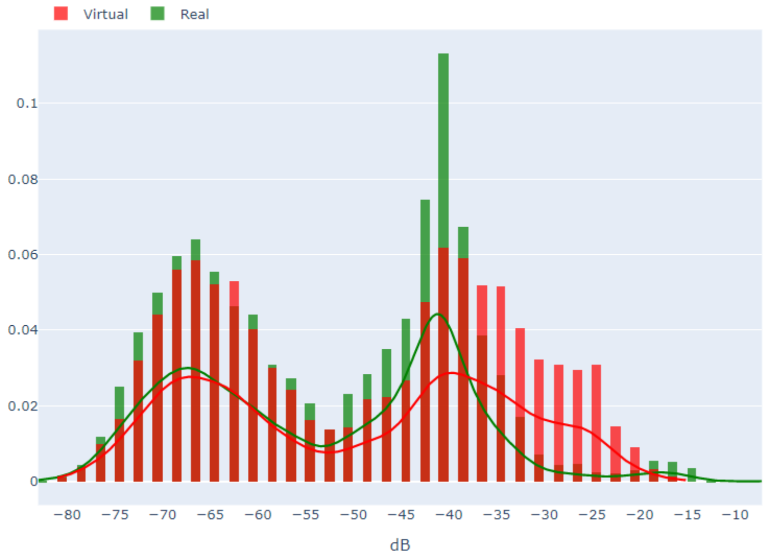


Figure A.11: Power distribution.

A.7.2 Dataset analysis

Table A.7 provides the major information of the datasets analysed in the following. Semi-anechoic dataset information is also provided for comparison purposes. To analyse and understand the efficiency of the algorithm proposed was necessary to have the ground truth on how many people were on-board and got on and off at each stop. To do this, a few days were spent manually counting people.. In A.7, it is also shown the key ground truth data: the number of devices in the semi-anechoic chamber which was constant over the observation window and the total number of people in the bus which was not always in the bus as they entered and exited the bus in different stops during the observation period.

Figure A.12 shows the distribution of the devices that have been observed among

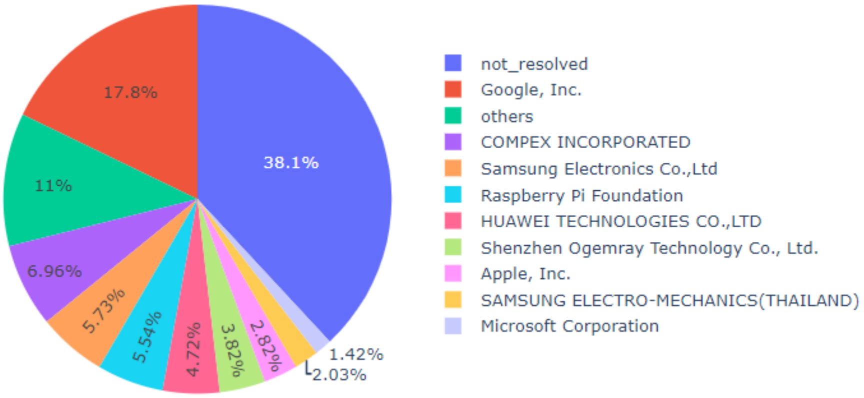


Figure A.12: Vendors distribution.

the various vendors. In “others” 84 vendors that appeared during tests with a presence lower than 1% are grouped. “not_resolved” represents the amount of traffic of which it was not possible to find the related vendor as some vendors perform randomisation in the OUI field as well making impossible their identification. By analysing the vendors’ identities it is easy to recognise that there are some that do not produce smartphones nor WiFi interfaces. This is the case of the category “Shenzhen Ogemray Technology Co., LTD”, which represents an IoT solutions manufacturer that does not produce WiFi interfaces for smartphones. This means that the sniffer is acquiring probes sent by domestic devices inside the buildings close to the street where the bus is passing by, as highlighted in Figure A.13. Indeed, this issue is particularly relevant when the bus stays at the bus stops as during this timeframe is more probable to capture stationary devices installed inside the buildings. For this reason, several processes to clean the data have been introduced, as explained in section A.5.

Splitting real and virtual MAC addresses and taking another look at the vendor distribution, the reality is more clear. All the virtual MAC addresses have Google as vendor id for 32.8% of the time, whereas for the other 67.2% the id is not resolvable. No other vendor is explicit in frames sent with a virtual MAC address which increases the difficulty in distinguishing devices.

Figure A.11 shows the power distribution for the captured frames, where it is possible to observe a double Gaussian-like distribution centred at two different power levels. It is obvious that one represents the devices within the bus, whereas the others those that are outside. This clear distinction among the two groups of devices suggested to introduce a power threshold to filter out the frames sent by devices which were inside, or outside the bus. Accordingly, this threshold was empirically determined by observing the power of frames generated by devices with known real MACs and varying the distance from the sniffer. The identified threshold that maximise the probability to identify the devices that were on-board was -53dB. It is important to highlight that this works with this specific setting and a model that could work in different settings must be devised.

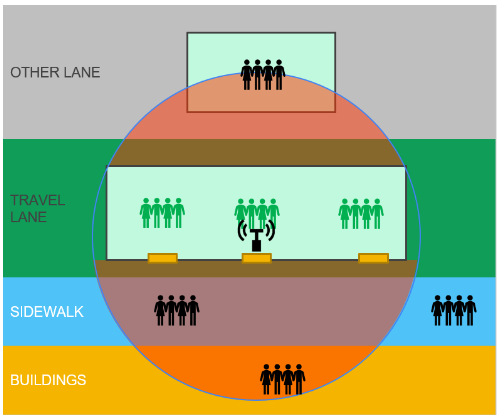


Figure A.13: Bus acquisition and data cleaning problem.

A.7.3 Performance analysis

Figure A.14 shows the number of passengers counted with the proposed algorithm on 4 sessions for the 4 different lines described above. The results of the counting performed takes into account the whole acquisition for each line and computes the number of devices at every bus stop. The graphs show the performance when using the three different clustering algorithms to analyse the impact of the different strategies. In particular, the features taken into account represent devices with constant density clusters; the DBSCAN clustering algorithm almost always provides more accurate estimates for the devices onboard the bus. A particular feature are the spikes in the estimation, due to people waiting out of the bus near the stop, which could not be excluded by the proposed algorithm. The OPTICS and the HDBSCAN algorithms are more sensitive to this issue.

In heavy congestion, but also when probe requests come from passengers waiting

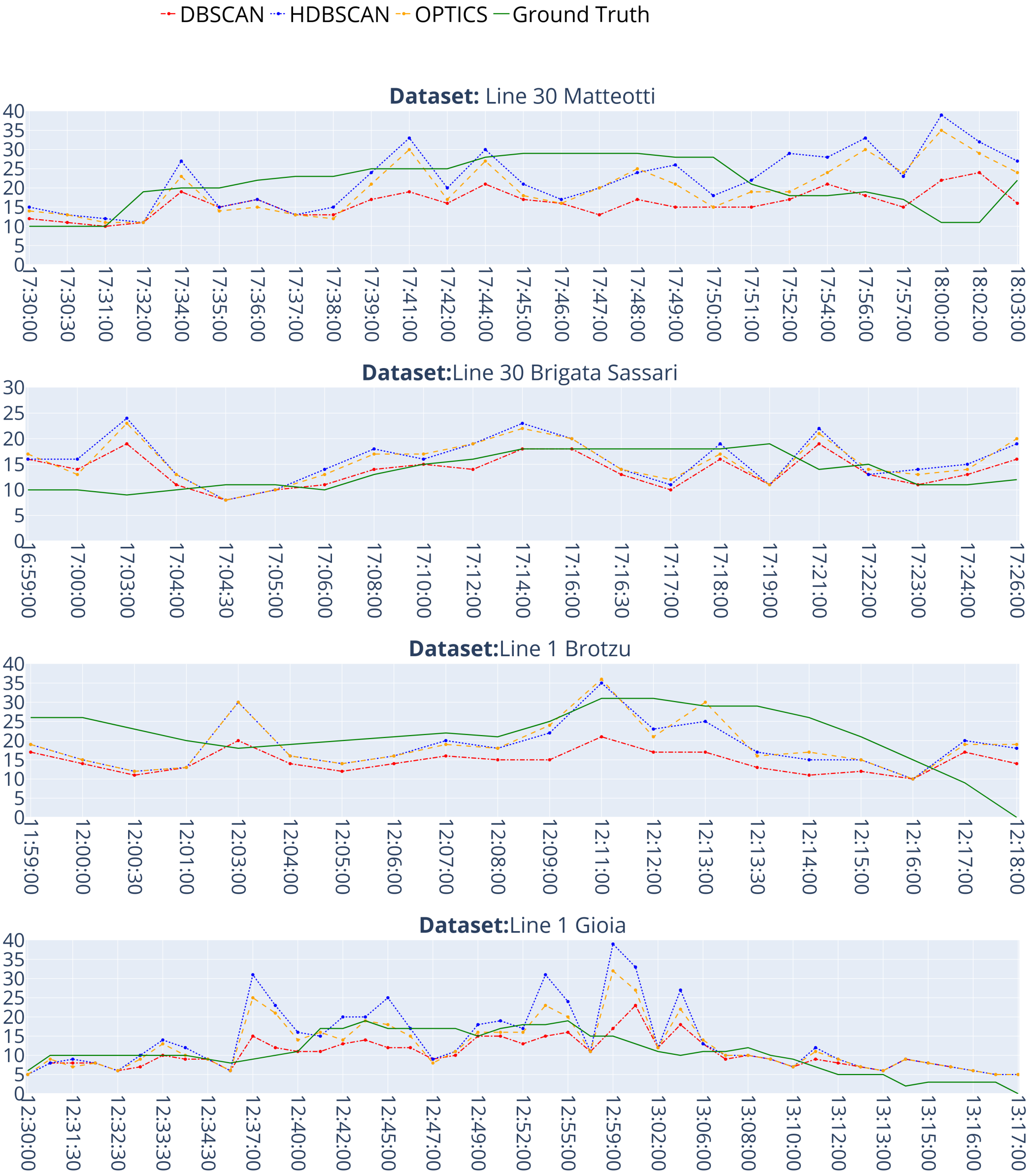


Figure A.14: Number of devices estimated for each dataset in the bus scenario.

for other buses near the stops (e.g. squares or stations), this generates noise in the features space with sporadic and rarefied points that the OPTICS and HDBSCAN algorithms mistake for low-density clusters; *i.e.* more devices. The issue could be easily resolved removing all the packets captured when the bus' doors are open. However, this cannot be implemented at the moment as the ground truth data does not include the information on the length of each bus stop. Table A.8 summarises the results with the relative error (average and standard deviation). As already highlighted the DBSCAN algorithm provides better results, with an accuracy as high as 74.25%.

Table A.8: Estimation results errors.

Scenario	**Dataset**	**HDBSCAN**	**DBSCAN**	**OPTICS**
Lab	Anechoic Chamber	3%	3%	3%
Real	Line 1 Brotzu	40%	31%	29%
	Line 1 Gioia	24%	27%	26%
	Line 30 B.Sassari	13%	27%	30%
	Line 30 Matteotti	29%	32%	32%
	Average	**26.5%**	**25.75%**	**29.25%**

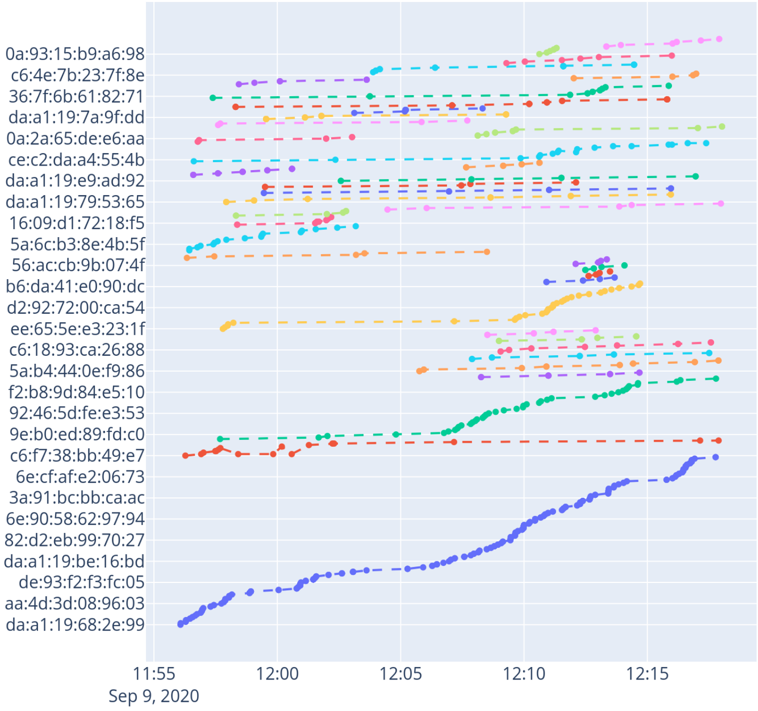


Figure A.15: Device traces with a pool of used MAC addresses (Dataset Brotzu).

Finally, Figure A.15 shows the several devices traces and the pool of virtual MAC addresses which have changed over time. It is a graphical result of the algorithm, through which it is quite simple to count the number of unique devices that have appeared in the T^W, simply by counting the number of coloured lines. This is an important result because it allows you to do a temporary tracking (limited to a few hours time) of the devices that use randomisation. Furthermore, by applying the algorithm to the urban public transport scenario, it is possible to easily create Origin-Destination matrices and automatically derive the demand of mobility in order to understand how citizens use public transport.

A.8 Conclusions

A novel de-randomisation algorithm has been presented, which relies on clustering Probe Request frames by considering the content and the rate at which the frames are emitted. The algorithm has been tested in a controlled environment, where only the frames generated by the test devices have been captured bringing to an accuracy of almost 97% when the observation window is at least 22 min long. This result represents an excellent basis to build solutions that can work in real scenarios. For this reason, further development and testing of a solution for an Automatic Passenger Counting system to be used on-board of buses was carried out. In this scenario, the average accuracy has been as high as 75%.

Bibliography

[1] S. Vollset, E. Goren, C.-W. Yuan, J. Cao, A. E. Smith, T. Hsiao, C. Bisignano, G. Azhar, E. Castro, J. Chalek, A. Dolgert, T. D. Frank, K. Fukutaki, S. Hay, R. Lozano, A. Mokdad, V. Nandakumar, M. Pierce, M. A. Pletcher, T. Robalik, K. M. Steuben, H. Y. Wunrow, B. S. Zlavog, C. Murray, Fertility, mortality, migration, and population scenarios for 195 countries and territories from 2017 to 2100: a forecasting analysis for the global burden of disease study, Lancet (London, England) 396 (2020) 1285–1306.

[2] 75% of the European population already lives in urban areas, says European Environment Agency, https://www.eea.europa.eu/themes/sustainability-transitions/urban-environment/urban-sustainability, [Online; accessed 20-Sept-2021] (2021).

[3] 68% of the World Population Projected to Live in Urban Areas by 2050, Says UN, https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html, [Online; accessed 20-Oct-2019] (2019).

[4] What are smart cities by European Commission, https://ec.europa.eu/info/eu-regional-and-urban-development/topics/cities-and-urban-development/city-initiatives/smart-cities_en, [Online; accessed 20-Sept-2021] (2021).

[5] Y. Feng, D. Duives, W. Daamen, S. Hoogendoorn, Data collection methods for studying pedestrian behaviour: A systematic review, Building

and Environment 187 (2021) 107329. doi:https://doi.org/10.1016/j.buildenv.2020.107329.

[6] General Data Protection Regulation (GDPR), https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en, [Online; accessed 07-Oct-2021] (2021).

[7] **Enrico Ferrara**, A. Liotta, M. Ndubuaku, L. Erhan, D. Giusto, M. Richardson, D. Sheffield, K. McEwan, A demographic analysis of urban nature utilization, in: 2018 10th Computer Science and Electronic Engineering (CEEC), 2018, pp. 136–141. doi:10.1109/CEEC.2018.8674206.

[8] **Enrico Ferrara**, A. Liotta, L. Erhan, M. Ndubuaku, D. Giusto, M. Richardson, D. Sheffield, K. McEwan, A pilot study mapping citizens' interaction with urban nature, in: 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech), 2018, pp. 836–841. doi:10.1109/DASC/PiCom/DataCom/CyberSciTec.2018.00-21.

[9] L. Erhan, M. Ndubuaku, **Enrico Ferrara**, M. Richardson, D. Sheffield, F. J. Ferguson, P. Brindley, A. Liotta, Analyzing objective and subjective data in social sciences: Implications for smart cities, IEEE Access 7 (2019) 19890–19906.

[10] **Enrico Ferrara**, L. Fragale, G. Fortino, W. Song, C. Perra, M. Di Mauro, A. Liotta, An ai approach to collecting and analyzing human interactions with urban environments, IEEE Access 7 (2019) 141476–141486. doi:10.1109/ACCESS.2019.2943845.

[11] M. Uras, R. Cossu, **Enrico Ferrara**, A. Liotta, L. Atzori, Pma: A real-world system for people mobility monitoring and analysis based on wi-fi probes, Journal of Cleaner Production 270 (2020) 122084. doi:https://doi.org/10.1016/j.jclepro.2020.122084.

[12] **Enrico Ferrara**, M. Uras, L. Atzori, O. Bagdasar, A. Liotta, Mobility analysis during the 2020 pandemic in a touristic city: the case of cagliari, accepted at 2021 IEEE IoT Vertical and Topical Summit for Tourism (IoT-VTST'21), September 2021 (2021).

[13] M. Uras, R. Cossu, **Enrico Ferrara**, O. Bagdasar, A. Liotta, L. Atzori, Wifi probes sniffing: an artificial intelligence based approach for mac addresses de-randomization, in: 2020 IEEE 25th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), IEEE, 2020, pp. 1–6.

[14] Improving wellbeing through urban nature (iwun), http://iwun.uk/, accessed: 20 August (2021).

[15] N. Zhong, J. Ma, R. Huang, J. Liu, Y. Yao, Y. Zhang, J. Chen, Research challenges and perspectives on wisdom web of things (w2t), The Journal of Supercomputing 64 (3) (2010) 862–882.

[16] B. Guo, D. Zhang, Z. Wang, Z. Yu, X. Zhou, Opportunistic iot: Exploring the harmonious interaction between human and the internet of things, Journal of Network and Computer Applications 36 (6) (2013) 1531–1539.

[17] A. Zanella, N. Bui, A. Castellani, L. Vangelista, M. Zorzi, Internet of things for smart cities, IEEE Internet of Things Journal 1 (1) (2014) 22–32. doi:10.1109/JIOT.2014.2306328.

[18] J. Jin, J. Gubbi, S. Marusic, M. Palaniswami, An information framework for creating a smart city through internet of things, IEEE Internet of Things Journal 1 (2) (2014) 112–121.

[19] J. Maas, R. A. Verheij, P. P. Groenewegen, S. De Vries, P. Spreeuwenberg, Green space, urbanity, and health: how strong is the relation?, Journal of epidemiology & community health 60 (7) (2006) 587–592.

[20] S. Ruiz-Correa, D. Santani, D. Gatica-Perez, The young and the city: Crowdsourcing urban awareness in a developing country, in: Proceedings of

the First International Conference on IoT in Urban Space, 2014, pp. 74–79.

[21] K. Samuelsson, M. Giusti, G. D. Peterson, A. Legeby, S. A. Brandt, S. Barthel, Impact of environment on peoples everyday experiences in stockholm, landscape and urban planning, vol. 171 (2018) 7–17.

[22] M. Richardson, D. Sheffield, Three good things in nature: noticing nearby nature brings sustained increases in connection with nature, Psyecology 8 (1) (2017) 1–32.

[23] T. Bakici, E. Almirall, J. Wareham, A smart city initiative: the case of barcelona, Journal of the Knowledge Economy 4 (2) (2012) 135–148.

[24] J. Lee, M. Hancock, M. Hu, Towards an effective framework for building smart cities: Lessons from seoul and san francisco, Technological Forecasting and Social Change 89 (2014) 80–99.

[25] Z. Khan, A. Anjum, K. Soomro, M. Tahir, Towards cloud based big data analytics for smart future cities, Journal of Cloud Computing 4 (2015) 1.

[26] M. Richardson, J. Hallam, R. Lumber, One thousand good things in nature: Aspects of nearby nature associated with improved connection to nature, Environmental Values 24 (5) (2015) 603–619.

[27] D. Strom, Big data makes things better, Dice Insights [Online] 17 (Dec 2018).

[28] H. Shahrokni, B. V. der Heijde, D. Lazarevic, N. Brandt, Big data gis analytics towards efficient waste management in stockholm, in: Proceedings of the 2014 conference ICT for Sustainability, Atlantis Press, 2014/08, pp. 140–147. doi:https://doi.org/10.2991/ict4s-14.2014.17.

[29] P. Anantharam, P. Barnaghi, K. Thirunarayan, A. Sheth, Extracting city traffic events from social streams, ACM Transactions on Intelligent Systems and Technology 6 (4) (2015) 1–27.

[30] R. Kitchin, Big data, new epistemologies and paradigm shifts, Big Data & Society 1 (2014) 1.

[31] Y. Fujiki, K. Kazakos, C. Puri, P. Buddharaju, I. Pavlidis, J. Levine, Neat-o-games, Computers in Entertainment 6 (2008) 2.

[32] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, C. Ratti, Real-time urban monitoring using cell phones: A case study in rome, IEEE Transactions on Intelligent Transportation Systems 12 (1) (2011) 141–151.

[33] Y. Qin, Q. Z. Sheng, N. J. G. Falkner, S. Dustdar, H. Wang, A. V. Vasilakos, When things matter: A survey on data-centric internet of things, Journal of Network and Computer Applications 64 (2016) 137–153.

[34] B. Guo, D. Zhang, Z. Wang, Z. Yu, X. Zhou, Opportunistic iot: Exploring the harmonious interaction between human and the internet of things, Journal of Network and Computer Applications 36 (6) (2013) 1531–1539.

[35] P. de Meo, E. Ferrara, F. Abel, L. Aroyo, G.-J. Houben, Analyzing user behavior across social sharing environments, ACM Transactions on Intelligent Systems and Technology 5 (1) (2013) 1–31.

[36] A. Sheth, Citizen sensing, social signals, and enriching human experience, IEEE Internet Computing 13 (4) (2009) 87–92.

[37] G. MacKerron, S. Mourato, Happiness is greater in natural environments, Global Environmental Change 23 (5) (2013) 992–1000.

[38] G. Miller, The smartphone psychology manifesto, Perspectives on Psychological Science 7 (3) (2012) 221–237.

[39] I. Bakolis, R. Hammoud, M. Smythe, J. Gibbons, N. Davidson, S. Tognin, A. Mechelli, Urban mind: Using smartphone technologies to investigate the impact of nature on mental wellbeing in real time, Biological Psychiatry 83 (2018) 9.

[40] K. Krippendorff, Content Analysis: An Introduction to its Methodology, SAGE Publications, Inc, Beverly Hills, 1980.

[41] R. Weber, Basic Content Analysis, SAGE Publications, Inc, 1990.

[42] A. K. Gopalakrishna, T. Ozcelebi, J. J. Lukkien, A. Liotta, Evaluating machine learning algorithms for applications with humans in the loop, in: 2017 IEEE 14th International Conference on Networking, Sensing and Control (ICNSC), 2017, pp. 459–464. doi:10.1109/ICNSC.2017.8000136.

[43] Google cloud platform - cloud vision api, https://cloud.google.com/vision/, accessed: 23 August (2021).

[44] K. Kalyanarangan, Text-clustering-api, GitHub repository [Online] 17 (Dec 2017).

[45] T. Panagopoulos, J. A. G. Duque, M. B. Dan, Urban planning with respect to environmental quality and human well-being, Environmental Pollution 208 (2016) 137–144, special Issue: Urban Health and Wellbeing. doi:https://doi.org/10.1016/j.envpol.2015.07.038.

[46] C. Yin, Z. Xiong, H. Chen, J. Wang, D. Cooper, B. T. David, A literature survey on smart cities, Science China Information Sciences 58 (10) (Oct. 2015). doi:10.1007/s11432-015-5397-4.

[47] J. Jin, J. Gubbi, S. Marusic, M. Palaniswami, An information framework for creating a smart city through internet of things, Internet of Things Journal, IEEE 1 (2014) 112–121. doi:10.1109/JIOT.2013.2296516.

[48] T. Bakıcı, E. Almirall, J. Wareham, A smart city initiative: the case of barcelona, Journal of the Knowledge Economy 4 (2) (2013) 135–148.

[49] N. Zhong, J. H. Ma, R. H. Huang, J. M. Liu, Y. Y. Yao, Y. X. Zhang, J. H. Chen, Research challenges and perspectives on wisdom web of things (w2t), The Journal of Supercomputing 64 (3) (2013) 862–882.

[50] S. Ghose, J. J. Barua, Toward the implementation of a topic specific dialogue based natural language chatbot as an undergraduate advisor, in: 2013 International Conference on Informatics, Electronics and Vision (ICIEV), 2013, pp. 1–5. doi:10.1109/ICIEV.2013.6572650.

[51] N. Svetlana, F. Daniel, B. Marcos, F. Casati, K. Georgy, Crowdsourcing for reminiscence chatbot design, in: Conference on Human Computation and Crowdsourcing (HCOMP 2018), 2018, pp. 1–5.

[52] L. Ni, C. Lu, N. Liu, J. Liu, Mandy: Towards a smart primary care chatbot application, in: International Symposium on Knowledge and Systems Sciences, Springer, 2017, pp. 38–52.

[53] G. Mackerron, S. Mourato, Happiness is greater in natural environments, Global Environmental Change 23 (2013) 992–1000. doi:10.1016/j.gloenvcha.2013.03.010.

[54] The smart app Shmapped, http://iwun.uk/shmapped/, accessed: 20 August (2021).

[55] A. Augello, M. Gentile, L. Weideveld, F. Dignum, A model of a social chatbot, in: G. D. Pietro, L. Gallo, R. J. Howlett, L. C. Jain (Eds.), Intelligent Interactive Multimedia Systems and Services 2016, Springer International Publishing, Cham, 2016, pp. 637–647.

[56] O. Bates, A. Friday, Beyond data in the smart city: learning from a case study of re-purposing existing campus iot, IEEE Pervasive Computing 16 (2) (2017) 54–60. doi:10.1109/MPRV.2017.30.

[57] D. Puiu, et al., Citypulse: Large scale data analytics framework for smart cities, IEEE Access 4 (2016) 1086–1108. doi:10.1109/ACCESS.2016.2541999.

[58] K. K. Mohbey, An efficient framework for smart city using big data technologies and internet of things, in: Progress in Advanced Computing and Intelligent Engineering, Springer, 2019, pp. 319–328.

[59] S. K. Datta, R. P. Ferreira da Costa, C. Bonnet, J. Harri, onem2m architecture based iot framework for mobile crowd sensing in smart cities, in: 2016 European Conference on Networks and Communications (EuCNC), 2016, pp. 168–173.

[60] J. M. Gutierrez, M. Jensen, M. Henius, T. Riaz, Smart waste collection system based on location intelligence, Procedia Computer Science 61 (2015) 120–127.

[61] Y. Li, W. Dai, Z. Ming, M. Qiu, Privacy protection for preventing data over-collection in smart city, IEEE Transactions on Computers 65 (5) (2016) 1339–1350.

[62] L. Van Zoonen, Privacy concerns in smart cities, Government Information Quarterly 33 (3) (2016) 472–480.

[63] M. Uras, R. Cossu, L. Atzori, Pma: a solution for people mobility monitoring and analysis based on wifi probes, in: 2019 4th International Conference on Smart and Sustainable Technologies (SpliTech), IEEE, 2019, pp. 1–6.

[64] S. Vaidya, P. Ambad, S. Bhosle, Industry 4.0–a glimpse, Procedia Manufacturing 20 (2018) 233–238.

[65] D. Toppeta, The smart city vision: how innovation and ict can build smart, livable, sustainable cities, Innov. Knowl. Found. (2010) 1–9.

[66] M. G. Demissie, S. Phithakkitnukoon, T. Sukhvibul, F. Antunes, R. Gomes, C. Bento, Inferring passenger travel demand to improve urban mobility in developing countries using cell phone data: a case study of senegal, IEEE Transactions on Intelligent Transportation Systems 17 (9) (2016) 2466–2478.

[67] D. Karamshuk, C. Boldrini, M. Conti, A. Passarella, Human mobility models for opportunistic networks, IEEE Communications Magazine 49 (12) (2011) 157–165.

[68] B. Guo, Z. Wang, Z. Yu, Y. Wang, N. Y. Yen, R. Huang, X. Zhou, Mobile crowd sensing and computing: The review of an emerging human-powered sensing paradigm, ACM Computing Surveys (CSUR) 48 (1) (2015) 7.

[69] P. Bahl, V. Padmanabhan, Radar: an in-building rf-based user location and tracking system, in: Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (Cat. No.00CH37064), Vol. 2, 2000, pp. 775–784 vol.2. doi:10.1109/INFCOM.2000.832252.

[70] M. Kotaru, K. Joshi, D. Bharadia, S. Katti, Spotfi: Decimeter level localization using wifi, SIGCOMM Comput. Commun. Rev. 45 (4) (2015) 269–282. doi:10.1145/2829988.2787487.

[71] A. Dagelić, T. Perković, M. Čagalj, Location privacy and changes in wifi probe request based connection protocols usage through years, in: 2019 4th International Conference on Smart and Sustainable Technologies (SpliTech), IEEE, 2019, pp. 1–5.

[72] E. Vattapparamban, B. S. Çiftler, I. Güvenç, K. Akkaya, A. Kadri, Indoor occupancy tracking in smart buildings using passive sniffing of probe requests, in: 2016 IEEE International Conference on Communications Workshops (ICC), IEEE, 2016, pp. 38–44.

[73] A. E. Redondi, M. Cesana, Building up knowledge through passive wifi probes, Computer Communications 117 (2018) 1–12.

[74] W. Wang, J. Chen, T. Hong, N. Zhu, Occupancy prediction through markov based feedback recurrent neural network (m-frnn) algorithm with wifi probe technology, Building and Environment 138 (2018) 160–170.

[75] A. Di Luzio, A. Mei, J. Stefa, Mind your probes: De-anonymization of large crowds through smartphone wifi probe requests, in: IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications, IEEE, 2016, pp. 1–9.

[76] F. Shi, K. Chetty, S. Julier, Passive activity classification using just wifi probe response signals, in: 2019 IEEE Radar Conference (RadarConf), 2019, pp. 1–6. doi:10.1109/RADAR.2019.8835660.

[77] E. Martin, O. Vinyals, G. Friedland, R. Bajcsy, Precise indoor localization using smart phones, in: Proceedings of the 18th ACM international conference on Multimedia, ACM, 2010, pp. 787–790.

[78] Z. Xu, K. Sandrasegaran, X. Kong, X. Zhu, B. Hu, J. Zhao, C. Lin, Pedestrain monitoring system using wi-fi technology and rssi based localization, International Journal of Wireless & Mobile Networks (2013).

[79] L. Schauer, M. Werner, P. Marcus, Estimating crowd densities and pedestrian flows using wi-fi and bluetooth, in: MOBIQUITOUS '14 Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, IEEE, 2014, pp. 171–177.

[80] F. Potortì, A. Crivello, M. Girolami, E. Traficante, P. Barsocchi, Wi-fi probes as digital crumbs for crowd localisation, in: 2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN), IEEE, 2016, pp. 1–8.

[81] A. Lesani, L. Miranda-Moreno, Development and testing of a real-time wifi-bluetooth system for pedestrian network monitoring, classification, and data extrapolation, IEEE Transactions on Intelligent Transportation Systems 20 (4) (2018) 1484–1496.

[82] C. Chilipirea, C. Dobre, M. Baratchi, M. van Steen, Identifying movements in noisy crowd analytics data, in: 2018 19th IEEE International Conference on Mobile Data Management (MDM), IEEE, 2018, pp. 161–166.

[83] M. W. Traunmueller, N. Johnson, A. Malik, C. E. Kontokosta, Digital footprints: Using wifi probe and locational data to analyze human mobility trajectories in cities, Computers, Environment and Urban Systems 72 (2018) 4–12.

[84] J. Andión, J. M. Navarro, G. López, M. Álvarez-Campana, J. C. Dueñas, Smart behavioral analytics over a low-cost iot wi-fi tracking real deployment, Wireless Communications and Mobile Computing 2018 (2018).

[85] F. Potortì, A. Crivello, M. Girolami, P. Barsocchi, E. Traficante, Localising crowds through wi-fi probes, Ad Hoc Networks 75 (2018) 87–97.

[86] A.-C. Petre, C. Chilipirea, M. Baratchi, C. Dobre, M. van Steen, Wifi tracking of pedestrian behavior, in: Smart Sensors Networks, Elsevier, 2017, pp. 309–337.

[87] P. Galluzzi, E. Longo, A. E. C. Redondi, M. Cesana, Occupancy estimation using low-cost wi-fi sniffers, ArXiv abs/1905.06809 (2019).

[88] W. Xi, J. Zhao, X.-Y. Li, K. Zhao, S. Tang, X. Liu, Z. Jiang, Electronic frog eye: Counting crowd using wifi, in: IEEE INFOCOM 2014-IEEE Conference on Computer Communications, IEEE, 2014, pp. 361–369.

[89] A. Kurkcu, K. Ozbay, Estimating pedestrian densities, wait times, and flows with wi-fi and bluetooth sensors, Transportation Research Record 2644 (1) (2017) 72–82. arXiv:https://doi.org/10.3141/2644-09, doi:10.3141/2644-09.

[90] E. A. Aronson, Location errors in time of arrival (toa) and time difference of arrival (tdoa) systems, Tech. rep., Sandia Labs., Albuquerque, N. Mex.(USA) (1977).

[91] G. Shen, R. Zetik, R. S. Thoma, Performance comparison of toa and tdoa based location estimation algorithms in los environment, in: 2008 5th Workshop on Positioning, Navigation and Communication, IEEE, 2008, pp. 71–78.

[92] J. Veselỳ, P. Hubáček, The tdoa system topology optimization from signal source position error estimation point of view, WSEAS Advances in Sensors, Signals and Materials (2010) 65–68.

[93] T.-K. Le, N. Ono, Refinement of time-difference-of-arrival measurements via rank properties in two-dimensional space, in: 2017 25th European Signal Processing Conference (EUSIPCO), IEEE, 2017, pp. 1971–1975.

[94] M. Wardman, A comparison of revealed preference and stated preference models of travel behaviour, Journal of transport economics and policy 22 (1) (1988) 71–91.

[95] B. Bonné, A. Barzan, P. Quax, W. Lamotte, Wifipi: Involuntary tracking of visitors at mass events, in: 2013 IEEE 14th International Symposium on" A World of Wireless, Mobile and Multimedia Networks"(WoWMoM), IEEE, 2013, pp. 1–6.

[96] A. J. Ruiz-Ruiz, H. Blunck, T. S. Prentow, A. Stisen, M. B. Kjærgaard, Analysis methods for extracting knowledge from large-scale wifi monitoring to inform building facility planning, in: 2014 IEEE International Conference on Pervasive Computing and Communications (PerCom), IEEE, 2014, pp. 130–138.

[97] H.-Y. Hsieh, S. W. Prakosa, J.-S. Leu, Towards the implementation of recurrent neural network schemes for wifi fingerprint-based indoor positioning, in: 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall), IEEE, 2018, pp. 1–5.

[98] M. Passafiume, S. Maddio, M. Lucarelli, A. Cidronali, An enhanced triangulation algorithm for a distributed rssi-doa positioning system, in: 2016 European Radar Conference (EuRAD), IEEE, 2016, pp. 185–188.

[99] S. Barai, D. Biswas, B. Sau, Estimate distance measurement using nodemcu esp8266 based on rssi technique, in: 2017 IEEE Conference on Antenna Measurements & Applications (CAMA), IEEE, 2017, pp. 170–173.

[100] C.-L. Yang, S. Bagchi, W. J. Chappell, Location tracking with directional antennas in wireless sensor networks, in: Microwave Symposium Digest, 2005 IEEE MTT-S International, 2005, pp. 131–134.

[101] S. Knauth, Study and evaluation of selected rssi-based positioning algorithms, in: Geographical and Fingerprinting Data to Create Systems for Indoor Positioning and Indoor/Outdoor Navigation, Elsevier, 2019, pp. 147–167.

[102] H. T. Friis, A note on a simple transmission formula, Proceedings of the IRE 34 (5) (1946) 254–256.

[103] K. W. Cheung, H.-C. So, W.-K. Ma, Y.-T. Chan, Least squares algorithms for time-of-arrival-based mobile location, IEEE Transactions on Signal Processing 52 (4) (2004) 1121–1130.

[104] W. Chen, J. Sun, L. Zhang, X. Liu, L. Hong, An implementation of ieee 1588 protocol for ieee 802.11 wlan, Wireless Networks 21 (08 2015). doi:10.1007/s11276-015-0898-z.

[105] F. Aletta, S. Brinchi, S. Carrese, A. Gemma, C. Guattari, L. Mannini, S. M. Patella, Analysing urban traffic volumes and mapping noise emissions in rome (italy) in the context of containment measures for the covid-19 disease, Noise Mapping 7 (1) (2020) 114–122. doi:doi:10.1515/noise-2020-0010.

[106] V. Harantová, A. Hájnik, A. Kalašová, Comparison of the flow rate and speed of vehicles on a representative road section before and after the implementation of measures in connection with covid-19, Sustainability 12 (17) (2020). doi:10.3390/su12177216.

[107] J. De Vos, The effect of covid-19 and subsequent social distancing on travel behavior, Transportation Research Interdisciplinary Perspectives 5 (2020) 100121. doi:https://doi.org/10.1016/j.trip.2020.100121.

[108] S. Wang, K. Wei, L. Lin, W. Li, Spatial-temporal analysis of covid-19's impact on human mobility: the case of the united states (2020). arXiv:2010.03707.

[109] C. Katrakazas, E. Michelaraki, M. Sekadakis, G. Yannis, A descriptive analysis of the effect of the covid-19 pandemic on driving behavior and road safety, Transportation Research Interdisciplinary Perspectives 7 (2020) 100186. doi:https://doi.org/10.1016/j.trip.2020.100186.

[110] G. Dantas, B. Siciliano, B. B. França, C. M. da Silva, G. Arbilla, The impact of covid-19 partial lockdown on the air quality of the city of rio de janeiro, brazil, Science of The Total Environment 729 (2020) 139085. doi:https://doi.org/10.1016/j.scitotenv.2020.139085.

[111] S. Sharma, M. Zhang, Anshika, J. Gao, H. Zhang, S. H. Kota, Effect of restricted emissions during covid-19 on air quality in india, Science of The Total Environment 728 (2020) 138878. doi:https://doi.org/10.1016/j.scitotenv.2020.138878.

[112] A. Otmani, A. Benchrif, M. Tahri, M. Bounakhla, E. M. Chakir, M. El Bouch, M. Krombi, Impact of covid-19 lockdown on pm10, so2 and no2 concentrations in salé city (morocco), Science of The Total Environment 735 (2020) 139541. doi:https://doi.org/10.1016/j.scitotenv.2020.139541.

[113] A. Kerimray, N. Baimatova, O. P. Ibragimova, B. Bukenov, B. Kenessov, P. Plotitsyn, F. Karaca, Assessing air quality changes in large cities during covid-19 lockdowns: The impacts of traffic-free urban conditions in almaty, kazakhstan, Science of The Total Environment 730 (2020) 139179. doi:https://doi.org/10.1016/j.scitotenv.2020.139179.

[114] A. Tobías, C. Carnerero, C. Reche, J. Massagué, M. Via, M. C. Minguillón, A. Alastuey, X. Querol, Changes in air quality during the lockdown in barcelona (spain) one month into the sars-cov-2 epidemic, Science of The Total Environment 726 (2020) 138540. doi:https://doi.org/10.1016/j.scitotenv.2020.138540.

[115] J. M. Baldasano, Covid-19 lockdown effects on air quality by no2 in the cities of barcelona and madrid (spain), Science of The Total Environment

741 (2020) 140353. doi:https://doi.org/10.1016/j.scitotenv.2020.140353.

[116] W. H. Jieya Yang, Effects of population density and traffic flow on covid-19 disasters in florida, Advancements in civil engineering technology 4 (2) (2020). doi:10.31031/ACET.2020.04.000585.

[117] Y. Chen, Y. Wang, H. Wang, Z. Hu, L. Hua, Controlling urban traffic-one of the useful methods to ensure safety in wuhan based on covid-19 outbreak, Safety Science 131 (2020) 104938. doi:https://doi.org/10.1016/j.ssci.2020.104938.

[118] M. Uras, **Enrico Ferrara**, R. Cossu, A. Liotta, L. Atzori, Mac address de-randomization for wifi device counting: combining temporal and content-based signatures, under Review on Computer Networks Journal (2021).

[119] Randomisation implementation timeline, https://www.arubanetworks.com/assets/tg/TD_Mac-Address-Randomization.pdf, accessed: 26 October (2021).

[120] J. Franklin, D. McCoy, Passive data link layer 802.11 wireless device driver fingerprinting, in: 15th USENIX Security Symposium (USENIX Security 06), USENIX Association, Vancouver, B.C. Canada, 2006, pp. 167–178.

[121] M. Vanhoef, C. Matte, M. Cunche, L. S. Cardoso, F. Piessens, Why mac address randomization is not enough: An analysis of wi-fi network discovery mechanisms, in: Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security, ASIA CCS '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 413–424. doi:10.1145/2897845.2897883.

[122] C. Matte, M. Cunche, F. Rousseau, M. Vanhoef, Defeating mac address randomization through timing attacks, in: Proceedings of the 9th ACM Conference on Security & Privacy in Wireless and Mobile Networks, 2016, pp. 15–20.

[123] M. Nitti, F. Pinna, L. Pintor, V. Pilloni, B. Barabino, iabacus: A wi-fi-based automatic bus passenger counting system, Energies 13 (2020) 1446.

[124] N. Suraweera, A. Winter, J. Sorensen, S. Li, M. Johnson, I. B. Collings, S. V. Hanly, W. Ni, M. Hedley, Passive through-wall counting of people walking using wifi beamforming reports, IEEE Systems Journal (2020) 1–7.

[125] M. Ribeiro, N. Nunes, V. Nisi, J. Schöning, Passive wi-fi monitoring in the wild: a long-term study across multiple location typologies, Personal and Ubiquitous Computing (2020) 1–15.

[126] M. Ribeiro, B. Galvão, C. Prandi, N. Nunes, Passive wi-fi monitoring in public transport: A case study in the madeira island (2020). arXiv:2006.16083.

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

[128] V. Vishnumurthy, S. Chandrakumar, E. G. Sirer, Karma: A secure economic framework for peer-to-peer resource sharing (2003).

[129] T. V. Kumar, Smart living for smart cities, in: Smart Living for Smart Cities, Springer, 2020, pp. 3–70.

[130] G. J. W. Kathrine, C. W. Joseph, Attacks, vulnerabilities, and their countermeasures in wireless sensor networks, in: Deep Learning Strategies for Security Enhancement in Wireless Sensor Networks, IGI Global, 2020, pp. 134–154.

[131] V. Brik, S. Banerjee, M. Gruteser, S. Oh, Wireless device identification with radiometric signatures, in: Proceedings of the 14th ACM international conference on Mobile computing and networking, 2008, pp. 116–127.

[132] G. Bartlett, J. Heidemann, C. Papadopoulos, Understanding passive and active service discovery, in: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, 2007, pp. 57–70.

[133] I. Purushothaman, S. Roy, Fastscan: a handoff scheme for voice over ieee 802.11 wlans, Wireless Networks 16 (7) (2010) 2049–2063.

[134] M. Uras, R. Cossu, L. Atzori, Pma: a solution for people mobility monitoring and analysis based on wifi probes, in: 2019 4th International Conference on Smart and Sustainable Technologies (SpliTech), 2019, pp. 1–6.

[135] J. Platt, et al., Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, Advances in large margin classifiers 10 (3) (1999) 61–74.

[136] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96, AAAI Press, 1996, p. 226–231.

[137] R. J. Campello, D. Moulavi, J. Sander, Density-based clustering based on hierarchical density estimates, in: Pacific-Asia conference on knowledge discovery and data mining, Springer, 2013, pp. 160–172.

[138] E. Nasibov, An alternative fuzzy-hierarchical approach to cluster analysis, in: Proceedings 7th International Conference on Application of Fuzzy Systems and Soft Computing, Germany, 2006, pp. 113–123.

[139] M. Ankerst, M. M. Breunig, H.-P. Kriegel, J. Sander, Optics: Ordering points to identify the clustering structure, ACM Sigmod record 28 (2) (1999) 49–60.