Gonçalo Ferreira

WHERETOGO

IDENTIFICATION AND CLASSIFICATION OF PLACES OF INTEREST

UNIVERSIDADE D COIMBRA
1 2 9 0

1 2 9 0

UNIVERSIDADE D
COIMBRA

Gonçalo Francisco Ferreira

WHERETOGO

IDENTIFICATION AND CLASSIFICATION OF PLACES OF INTEREST USING ANONYMIZED MOBILE COMMUNICATION DATA

VOLUME 1

Dissertation in the context of the Master in Informatics Engineering, Specialization in Inteligent Systems advised by Professor Carlos Bento and Professor Ana Alves and presented to the Faculty of Sciences and Technology/Department of Informatics Engineering

June 2021

UNIVERSIDADE D
COIMBRA

Gonçalo Francisco Ferreira

WhereToGo

Identification and classification of places of interest using anonymized mobile communication data

Dissertation in the context of the Master in Informatics Engineering, Specialization in Inteligent Systems advised by Professor Carlos Bento and Professor Ana Alves and presented to the Faculty of Sciences and Technology /Department of Informatics Engineering.

June 2021

Abstract

The importance of understanding human mobility is transversal to multiple fields of study and practical applications, from ad-hoc networks to smart cities, from recommendation systems on social networks to transportation planning. Digital location traces can help extract insights about how citizens experience their cities. Moreover, these footprints serve as the foundation to detect patterns and offer personalized products and experiences to people.

In recent years Call Detail Records (CDRs) have shown great potential for research purposes on the analysis of movement patterns and identification of important places. This data type offers significant advantages over alternatives as it is ubiquitous, computationally inexpensive and easier to collect in large-scale. However, previous works using CDRs tend to focus on modeling spatial and temporal patterns of human mobility, not paying much attention to the semantics of places, thus failing to model people's motivation behind their mobility. Although the question of identifying user's activities is far more stabilised with detailed information types, such as GPS, we feel that there is still room for improvement and innovation working with CDRs.

In this work, we aim to create a system that can identify individual users routine places and attach some semantic meaning to them, inferring the motivations for day-to-day mobility. We investigate and apply methods to prepare the data, mitigating known flaws of CDRs, including the removal of anomalous records. We then find the spatio-temporal patterns of user's records using two clustering algorithm, the density-based DBSCAN, for home and workplace, and the partition based K-means clustering for other routine places. Finally we classify the discovered places according to most likely activity, using nearby Points of Interest.

This research also introduces a novel methodology to position user's localization from CDRs through improved understanding of antenna's signal areas. Moreover, several methods were developed and applied to validate some of the generated results using ground-truth data obtained from 4600 users.

Keywords — Call Detail Records, Clustering Algorithms, Human Mobility, Meaningful places, Mobile Phone Data, Points of Interest.

Resumo

A obtenção de conhecimento sobre a mobilidade humana tem uma importância transversal a diversos campos científicos e aplicações práticas, deste redes ad-hoc a cidades inteligentes, de sistemas de recomendação em redes sociais a soluções de planeamento de transportes. Registos de localização digitais permitem extrair o conhecimento de como as pessoas experienciam as cidades e o espaço em geral. Estas 'pegadas' de localização servem assim como base para detetar padrões de mobilidade, muitas vezes com o objetivo final de criar produtos e serviços personalizados.

Recentemente *Call Detail Records* (CDRs) demonstraram ter um grande potencial para pesquisas relacionadas com a análise de padrões de movimento individuais e identificação de locais importantes. Este tipo de dados oferece vantagens sobre as alternativas, uma vez que é ubíquo, computacionalmente menos exigente de avaliar e mais fácil de recolher em grande escala do que, por exemplo, *GPS traces*. Contudo, trabalhos anteriores que recorrem a CDRs tendem a focar-se em modelar apenas os padrões espaciais e temporais da mobilidade, não dando atenção à semântica dos locais, ignorando assim a procura pelas motivações por detrás da mobilidade. Embora a questão de inferir atividades realizadas esteja mais estabilizada com tipos de informação mais detalhados como o GPS, ainda existe muito espaço para melhorias e inovações utilizando CDRs.

Nesta tese pretendemos criar um sistema capaz de identificar locais de rotina de utilizadores e adicionar-lhes um significado semântico, procurando assim inferir as motivações para a mobilidade do dia-a-dia. Para chegar ao objetivo final começamos por investigar e aplicar métodos para preparar os dados, incluindo a remoção de registos anómalos, mitigando problemas conhecidos associados aos CDRs. De seguida, encontramos padrões espaciais e temporais nos registos utilizando dois algoritmos de clustering, o DBSCAN, para os locais de casa e trabalho, assim como o K-means, para as restantes localizações de rotina. Finalmente classificamos os locais de rotina encontrados de acordo com a atividade mais provável, para tal recorrendo a Pontos de Interesse na vizinhança.

Este trabalho também propõe um novo método para posicionar a localização de um utilizador nos CDRs com recurso a conhecimentos sobre a direção e alcance do sinal de antenas de telecomunicações. Adicionalmente foram aplicados métodos para validar alguns dos resultados obtidos no decorrer do projeto, usando dados *ground-truth* recolhidos de cerca de 4600 utilizadores.

Palavras-Chave — Algoritmos de Clustering, Call Detail Records, Dados de comunicações móveis, Locais importantes, Mobilidade Humana, Pontos de Interesse.

This page intentionally left blank.

Acknowledgements

I would like to thank the AmILab team at the The Centre for Informatics and Systems of the University of Coimbra, without whom I would not have been able to complete this thesis.

Thank you to my advisors and mentors, Professor Carlos Bento and Professor Ana Alves for their enthusiasm for the project, for their support, encouragement and patience. Special thanks to Professor Marco Veloso for his availability, providing valuable feedback at every step. Their invaluable advice, insight and knowledge into the subject matter steered me through this research.

I would like to thank all my friends and colleagues for a cherished time spent together, encouragement and support all through my studies. Finally, I would like to express my gratitude to my family and my girlfriend who endured this long process with me, always offering support and love.

This page intentionally left blank.

Contents

List of Figures

List of Tables

This page intentionally left blank.

Abbreviations

AmILab: Ambient Intelligence Laboratory

API: Application Programming Interface

CDRs: Call Detail Records

CISUC: Centre for Informatics and Systems of the University of Coimbra

DBSCAN: Density Based Spatial Clustering of Applications with Noise

Eps: Epsilon

GPS: Global Positioning System

GSM: Global System for Mobile Communications

MVPs: Most Visited Places

OVPs: Occasionally Visited Places

POIs: Points of Interest

QGIS: Quantum Geographic Information System

WGS84: World Geodetic System 1984

This page intentionally left blank.

Chapter 1

Introduction

1.1 Context and Motivation

Human mobility has become a prominent research field in recent years. There is a growing need to understand how people move and use the urban space in their daily routines. We know for a fact that human trajectories are characterized by a high degree of temporal and spatial regularity. For each individual there is frequent time-independent travel distance and a significant probability to return to a few highly frequented locations, with only few diversions of habitual trips to visit new locations [1]. These hidden patterns of motion have importance in applications such as urban planning, traffic forecasting and the spread of biological and mobile viruses. Ubiquitous computing has in fact unlocked the potential to determine personal movements of the masses that previously were only modelled using household surveys and national or regional census.

In today's society nothing could be more ubiquitous than mobile phones, each of them a potential sensor providing a constant data stream. On this basis, specialized spatio-temporal datasets like GPS records have shown enormous potential as a knowledge base about human mobility patterns. In spite of that, due to the overhead of collecting and analyzing such detailed and high frequency location logs at a large scale, many researchers have been urged to explore other data sources as potential proxies for human mobility.

With this in mind we take special interest in Call Detail Records. A Call Detail Record (CDR) is a form of data that documents the details of how a user is interacting with the cellular network. These records contain information fields like origin/destination tower ID, user ID, time of start and duration on any type of communication. Collecting the cell tower ID that the user is connected at the time of an event means that an accurate location is not possible, only an approximation, provided we obtain the Cartesian coordinates of the tower's position. In the case of CDRs records are not normally recorded in regular time intervals, unlike GPS, only when a network event occurs (e.g., a call is made). This leads to data that is both spatially sparse and temporally irregular.

Additionally, the lack of consent on the usage of this data leads to some privacy issues, a growing worry in our days. As this is a significant concern, telecommunication companies go to great lengths to ensure that data coming from the smartphone sensors or cellular networks cannot identify a user under any circumstance. Call Detail Records, that are generally already collected for billing purposes and network analysis, suffer an alteration, namely, the anonymisation of the user identity in such a way that his/her mobile phone number is encrypted.

Despite the significant benefits at scale of this information type, the detail in the locations recorded is a challenge for research efforts on the individual analysis of users. Given we only have the position of a cell tower antenna, which range of action spans at the minimum a few hundred meters and at the maximum several kilometers, it is very difficult to pinpoint an accurate user's location. This uncertainty in location, as expected makes it very challenging to understand and classify each user's important places. Although the question of identifying and classifying places is far more stabilized with detailed information types, such as GPS, we feel that there is still room for improvement and innovation working with Call Detail Records.

1.2 Objectives

The WhereToGo project goal is to develop a system that can identify user's regular places of stay and attach some semantic meaning to them, inferring the motivations behind day-to-day mobility. The focus is more on classifying activities (e.g. shopping, dinning, outdoor recreation, etc.) outside of the normal commute of home/work but that still have some sort of regularity. We will be using common data from an aggregation of telecommunication events (e.g. calls, messages, mobile internet connections, etc.) as the base for this project. This data was made available by one of the largest telecommunications service provider in Portugal who also is the main interested party, to which our results will directly feed back to.

Using several types of complementary records, including some new ones presented by the service provider, we hope to surpass some of the problems related to the sparse and irregular time intervals, as filling the gaps of mobility traces facilitates trajectory reconstruction and inferring people's routine places. In using both CDRs and Points of Interest (POIs) datasets to detect and classify important places, we gather a wealth of information that can be used for other studies and applications both with a larger or smaller scope. On the macro scale we obtain information on movement patterns and urban space usage by a large sample of the population, which is valuable in areas such as urban, transportation and business planning. On a small scale, by analyzing each user's favorite places we can infer about their habits and tastes, valuable parameters for recommendation systems and other marketing purposes. Ultimately the client and final recipient of this data can take advantage of this information and use it to adapt their marketing campaigns and improve quality of service by taking into account the client's new profiling.

As for the validation and evaluation of the created system and applied algorithms we also have to rely on the telecommunications company as the data source. We know that without proper ground truth data available we would only have the option to use dated population census to compare with our obtained results. This has been a major focus of our attention throughout the work period, actively working with the client in order to gather the required annotated data to serve as ground truth.

1.3 Internship

This internship was hosted by the Ambient Intelligence Laboratory - AmILab, part of the Centre for Informatics and Systems of the University of Coimbra (CISUC). AmILab work focuses on understanding and predicting relevant dynamics in the urban areas for urban planning and design. The group specializes on the integration and analysis of crowd-sourced, social media and open data, as well as stakeholder database, through AI techniques.

Early on it was mutually agreed that all the produced documentation would be written in English, including this report.

Weekly meetings, usually scheduled to Friday morning, took place with interns and advisors to make a weekly report about progress, the status of the individual projects, encountered challenges, and other issues that may have arisen throughout the previous days. A PowerPoint presentation was often brought by each intern to help in visualize and explain the progress made to the whole team.

In addition, every two weeks, a more formal meeting was held with the client and main data provider. These meetings aimed to show the developments of our research projects and the potential behind the obtained results. They also served the purpose of discussing the client's view on future steps and probe the availability of new and improved data.

All these factors contributed to us never loosing sight of the short and long term goals of the project and knowing what had to be done to reach said goals in their respective deadlines.

1.4 Expected Contributions

In the course of the work developed along the WhereToGo project the expected contributions are the following:

- Documentation of the current State of the Art regarding human mobility studies using CDRs.
- Development of a model to infer meaningful places plus their classification in terms of functionalities.
- Validation of the developed model using real ground truth data and analysis of the results obtained.

- Writing of a scientific paper.

1.5 Document Structure

This document is divided into seven chapters. The current chapter explains the motivations behind the project, important context information, gives an overview of the project, the potential risks, goals and expected contributions.

Chapter Two contains detailed explanations of the concepts needed to understand both State of the Art and the followed methodology for this thesis work.

Chapter Three comprises the State of the Art, where we look over the current best practices and implemented methods, where the technology is leading, what can be learned from each method and previous study.

Chapter Four is exclusively about the datasets used throughout this work. An explanation of the data tables and a brief exploratory analysis is present in this section.

Chapter Five provides an outline of the methodology. The research approach, data analysis techniques and description of the algorithms are present here.

In Chapter Six we report experiments conducted to test the behavior of the proposed methodology as well as describe and discuss the results obtained. Validation and Evaluation, when ground-truth data permited, is also included.

Finally, Chapter Seven addresses the conclusions and future work.

Chapter 2

Background Concepts

Starting off, a basic understanding of some methods or algorithms is needed to ensure full comprehension of the studies described in the next Chapter as well as the work carried out in Chapter 4. As to improve the flow of this paper and not repeat certain definitions all background information needed will be found here.

2.1 Cellular Networks

In this section we make some brief explanations of how the cellular network is organized, according to the best of our understandings. This was partially obtained through both literature, our work in these past semesters and our knowledge exchange with the telecommunication provider.

2.1.1 Base Stations and Antennas

Cellular network are communication networks with transceivers located at various fixed locations called base station. These base stations are distributed for propagating the signal to the mobile users. Because base stations are fixed and have a limited coverage area it is suggested to have a sufficient number to obtain the best possible coverage area, so that the mobile users ideally have equal signal levels at any point [2].

Generally base stations will have a tower with several antennas, each covering a certain angle in the surrounding area of the tower, as is visualized in Figure 2.1. The smallest element in the network is in fact the network antenna, to which the mobile devices are connected when they are subscribed to the network. The antenna is responsible for radio communications between the network and the mobile devices [3]. Each one has a defined a region of coverage, also known as "cell", and a unique identifier. The number of antennas will depend on the tower as there are different antennas for different frequencies and protocols. If a tower has several types of services (e.g. 2G, 3G and 4G), this will mean several dedicated antennas for each one. Also, antennas of dissimilar service type overlap each other in terms of region of coverage, as they have different effective ranges and frequencies of signal.

Figure 2.1: Visualization of a cellular tower and its cells.

2.1.2 Network Connection and Records

According to the authors in [4], there are two main kinds of connections types when it comes to mobile devices: event-driven and network-driven. The first one requires a user's participation in a deliberate network event (e.g. making a call, sending a message, accessing mobile internet, etc.). The second one does not require it. For these connections records are generated periodically without human intervention. In this work we will be using a mixture of both types in the hopes that the periodic records will serve to fill the temporal gaps of regular Call Detail Records.

In the topic of connections its important to explain that although a user connects to an individual antenna, with a specific region of coverage (cell), the location in the records that is attributed to that connection will always be the tower location. So in normal Call Detail Records(CDRs) if a user makes a events in different cells of the same tower the geographical coordinates associated with that event will be the same. This is true in most works encountered in literature because information about the individual cells are not usually disclosed or in any way publicly available.

Another relevant factor to take into consideration is that the cell the user is connected is not always the closest one to his/her geographical position. It depends on factors such as antenna signal strength, interference with landscape features, buildings and even network load. Furthermore, different kinds of mobile network events use different antennas depending on the needed service type. A user that stays in the same location might connect to different antennas to make a call using 2G or access mobile internet on 4G , provided that his position is inside the regions of coverage.

2.2 Clustering Methods

Data clustering is the process of identifying natural groupings or clusters within multidimensional data based on some similarity measure (e.g., Euclidean distance) [5]. A cluster is usually identified by a centroid and the quality of clustering depends on the similarity measure used and its implementation. An effective clustering technique will create clusters with high intra-class similarity and low inter-class similarity. It is a central process in development of systems related to pattern recognition, machine learning and Artificial Intelligence with algorithms used in many applications, such as image segmentation, vector and color image quantisation, data mining, compression, etc.

2.2.1 Types of Clustering

As there are hundreds of published algorithms, it is difficult to define strict categories that encompass all. The initial well adopted criterion suggested dividing the methods into two main groups: hierarchical and partitioning methods. More recently [6] suggests categorizing the methods into five main categories: hierarchical, partitioning, density-based, model-based and grid-based.

Hierarchical Clustering as the name suggests creates a group of nested clusters structured as a hierarchical tree. Hierarchical procedures can be either agglomerative, starting with each data point in its own cluster and merging the most similar pair of clusters successively to form a cluster hierarchy, or divisive, starting with all the data-points in one cluster and recursively dividing each cluster into smaller clusters.

The Partitioning algorithms, unlike hierarchical, result in a group of N clusters found simultaneously. Each item belongs to a unique cluster, meaning no overlapping occurs in the final groups. A cluster may be denoted by a centroid or a cluster representative, a sort of summary description of all the entities enclosed in a cluster. This class of algorithms can be distinguished by some characteristics. They usually divide the data into a predefined number of clusters, are generally iterative algorithms that converge to local optima and results differ depending on the selection of the initial starting partition. Most partitioning methods cluster objects based on the distance between them. Such methods can find only spherical-shaped clusters and encounter difficulties in discovering clusters of arbitrary shapes.

Density-based approaches have been, as the name implies, developed based on the notion of density. They assume that the points that belong to each cluster are drawn from a specific probability distribution. Their general idea is to continue growing a given cluster as long as the number of data points in the "neighborhood" exceeds some defined threshold. For example, for each data point within a given cluster, the neighborhood of a given radius has to contain at least a minimum number of points. Such a method can be used to filter out noise or outliers and discover clusters of arbitrary shape.

Grid-based methods quantise the object space into a finite number of cells that form a grid structure. All the clustering operations are performed on the grid structure (i.e., on the quantised space). The main advantage of this approach is its fast processing time, which is typically independent of the number of data objects and dependent only on the number of cells in each dimension in the quantised space.

Model-based methods attempt to optimize the fit between the given data and some mathematical models. Unlike conventional clustering, which identifies groups of objects, model-based clustering methods also find characteristic descriptions for each group, where each group represents a concept or class. The most frequently used induction methods are decision trees and neural networks.

Now after a brief overview of clustering methods we focus our attention more on two Partition and Density-based clustering algorithms. The next subsection conducts a deeper explanation of the two most relevant clustering algorithms in our state-of-the-art research, with both of them having been implemented in our work's methodology.

2.2.2 K-means Algorithm

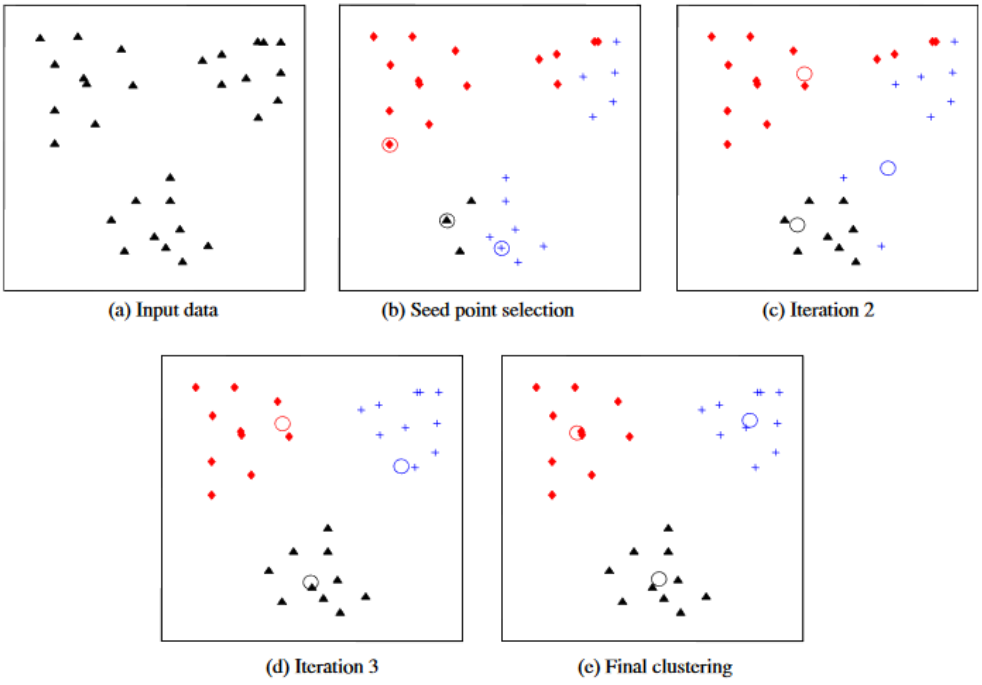


Figure 2.2: Illustration of K-means algorithm [7].

Probably the most commonly used partition algorithm by researchers is the iterative K-means approach [6]. This method requires as inputs three main parameters: the number of clusters K, cluster initialization and a distance metric. Normally initialization points will be chosen randomly and the distance metric used is Euclidean [7]. The important value of K will most likely depend on the application, its optimization being a simple process of running for different values and the partition that appears the most meaningful is selected. Initialization can also have a meaningful impact as it can lead to different final clusters. As such K-means should be run, for a given K, with multiple different initial partitions and the partition with the smallest squared error is chosen.

Initially, each data point is associated with one of the K clusters according to its distance to the centroids of each cluster. As can be seen from the example in Figure 2.2(b), where circles correspond to centroids and the remaining points have the same color and shape if the centroid that is closest to them is the same. Then, new centroids are calculated, and the classification of the data points is repeated for the new centroids, as indicated in Figure 2.2(c), where it is shown the new position of the centroids, different from the previous iteration. The process is repeated until no significant changes of the centroids positions are observed at each new step, as shown in Figure 2.2(e).

2.2.3 DBSCAN Algorithm

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) directly searches for connected dense regions in the feature space by estimating the density using the Parzen window method [8]. The performance of this algorithm depends on two parameters: neighborhood size in terms of distance (Eps), and the minimum number of points in a neighborhood for its inclusion in a cluster (MinPts).

Given a set D of points, the clustering task is therein reduced to using core points and their neighborhoods to form dense regions, where the dense regions are clusters. For a core point q and a point p, we say that p is directly density-reachable from q (with respect to Eps and MinPts) if p is within the Eps-neighborhood of q. Using the directly density-reachable relation, a core point can aggregate all objects from its Eps-neighborhood into a cluster.

The main functioning steps are as following: initially, all points in a given dataset D are marked as “unvisited”; DBSCAN randomly selects an unvisited point p; this point is marked as “visited,” and checked whether his Eps-neighborhood contains at least MinPts, as shown in Figure 2.3(a); if not, p is marked as a noise point; otherwise, a new cluster C is created for point p, and all the objects in the Eps-neighborhood are added to a candidate set, N.

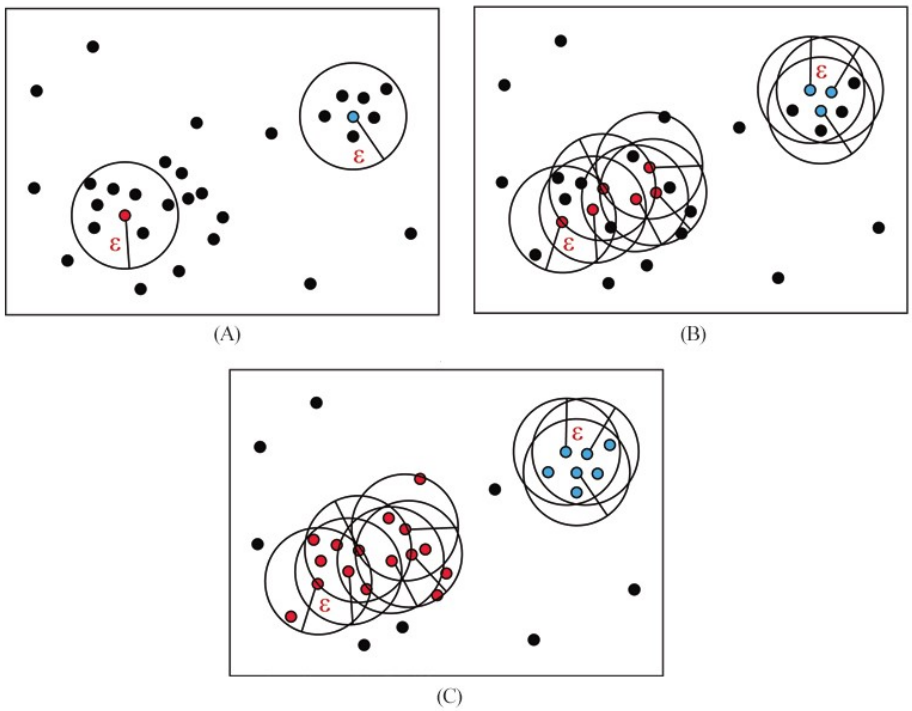


Figure 2.3: Illustration of the DBSCAN algorithm (Adapted from [9]).

DBSCAN iteratively adds to C those candidate points that do not belong to any cluster. In this process, for a point in the candidate set that carries the label "unvisited", DBSCAN marks it as "visited" and checks if its neighborhood has at least MinPts objects, as per Figure 2.3(b). Those objects in the Eps-neighborhood of the candidate point are added to N. DBSCAN continues adding objects to the cluster C until it can no longer be expanded, that is, N is empty, like in Figure 2.3(c). To find the next cluster, DBSCAN randomly selects an unvisited object from the remaining ones. The clustering process continues until all objects are visited.

Chapter 3

State of the Art

In this section it is given an up-to-date explanation of the most relevant state-of-the-art methods. The research stage throughout this thesis focused on Call Detail Record uses in mobility studies, finding meaningful places, making the distinction between routine and non-routine visits and classifying them for their purpose.

With the increasing popularity of personal mobile devices and location-based applications, large-scale trajectories of individuals are being recorded and accumulated at a faster rate than ever, which makes it possible to understand human mobility from a data-driven perspective. As a consequence of this ever-increasing availability of data, researches based on Spatial-temporal data have received a lot of interest, with a large spectrum of methods developed to analyze human mobility. In parallel, interest in human movement has shifted from raw movement data analysis to more application-oriented ways of analyzing segments of movement suitable for the specific purposes of the application. This trend has promoted semantically rich trajectories, rather than raw movement, as the core object of interest in mobility studies [10].

There are several technologies capable of capturing trajectories but their basic idea is the same, the ability to capture the movement of an object moving in geographical space over some period of time. Movement capture results in generating for each moving object its movement track. A movement track basically consists in the temporal sequence of the spatio-temporal positions, that is, a timestamp and point pair, recorded for the moving object. However, depending on the capabilities of the device and data collection type, additional data, for example, the instant speed or stillness, acceleration, direction, and rotation, may complement the (timestamp, point) pairs. We call raw data the data as captured from the device [11]).

3.1 Call Detail Records for Mobility Studies

There are some compelling reasons for using Call Detail Records (CDRs) versus more accurate location sources. First and foremost, compared to other location log data types, the overhead for collection and analysis is inferior. Also, data collection cannot be turned on or off by the user, it is generated by regular usage of a mobile communication device and stored by the mobile network provider. The fact that this data is already collected automatically by the cellular network to assist call reporting, billing, and to analyze existing infrastructure, means there is no need to conduct specific studies to gather it. Knowing this, potentially every phone in one provider's network can be used as a data source, resulting in enormous amounts of information regarding thousands or even millions of individuals that can be used for research purposes. Also, there is no cost on the user's side, no app has to be running on the device and no additional battery life is consumed.

Despite the listed benefits, the use of this type of data raises questions regarding the validity of the researches and the obtained conclusions, seeing that CDRs provide limited accuracy along both the spatial and temporal dimensions. In fact, studies like the one in [12], were conducted with the objective of proving that identifying users most significant locations, like home and work, is possible with a high degree of success. Another research work, around the same time, [13] compared CDR-based individual trajectories with reference information from public transport data, i.e., GPS logs of taxis and buses, as well as subway transit records. They found these two types of information to match with a good enough accuracy for extracting user's movements. The main conclusion from both these examples is that it is necessary to reduce the inherent localization error present in CDRs and other kinds of Global System for Mobile Communications (GSM) data records. GSM being a standard for all mobile communication encompassing all services between the mobile phone and the network. So, a starting point when working with CDRs can be by means of techniques to minimize the spatial uncertainty and temporal irregularity as well as detecting and treating networks event records that might create misleading information.

There are some exploratory attempts at filling the temporal gaps, 'completing' the records with ones generated by algorithms or fusing with other kinds of complementary data. An interesting comparative study in this area [14] proposed the approach of completing CDRs with several rule-based techniques, and validating the results with ground-truth GPS data. Their conclusions actually provide a first clear ranking for techniques for CDR data completion. Specifically, they show that a solution that extends for a limited amount of time the stays of users at known locations, and places users at their home locations, within a night time period, achieves improved accuracy and a fair coverage. A Previous proposal in the literature [15] included completing the records in a static way, only extending the user's position on the last known location until a new record appears. The downside of this particular work [14], is that they did not have access to a mobile operator CDRs, as these are not easily available online due to the high privacy concerns with this information type. Faced with this problem they choose to down-sample their ground-truth GPS dataset to approximate CDR records, as such we are left with an incomplete validation.

When it comes to integrating different types of data collected from mobile devices to discover human mobility patterns there are also a few examples. Montoliu et al.,[16] exemplifies the discovery of human places-of-interest from a multitude of mobile phone data, such as GPS, Wi-Fi, GSM, sensors, etc., with relative success. Han et al.,[6] extracted users semantic features from mobile phone trace data, POIs and real estate price data. Zhang et al.,[17] studied the user spatial behaviors and trade area analysis through geo-tagged tweets, Foursquare check-ins and CDRs.

As previously stated, preserving privacy is always a concern, working with data that can be considered sensitive personal information. To that effect, the work presented in [18] demonstrated that, with certain protection techniques applied, re-identification of an individual via frequently visit locations, co-location pairs or spatial temporal data points with a high probability is not possible. Nevertheless, as CDRs are still gathered and processed remotely at the end of the day, it remains for end users a question of trust. Some new attempts are now exploring the possibility to perform mobility modelling tasks and create recommendation systems locally on a user device [19]. This new paradigm would work around the challenges raised by the development of an ambient and context-aware personal recommender system in a strong privacy environment.

3.1.1 Load Sharing

Working with Call Detail Records provides several challenges, some due to the way the network operates and the way the locations are collected. One potential issue identified by the state-of-the-art literature is Load Sharing. On account of variations in the cellular traffic at certain areas or periods in time, the user might see a temporary shift of connection to different tower, giving an apparent impossible shift in the user's position, as per Figure 3.1.

This phenomenon is often called Load Sharing [20] and consists of the network transferring traffic from overloaded cell towers to neighboring ones, as a way to provide service for all clients. These anomalous records, although occasional for a dataset consisting of a large number of records, can affect studies regarding the user's mobility, giving incorrect information to the learning algorithms. Therefore, it is better if they are ignored or removed from the dataset.

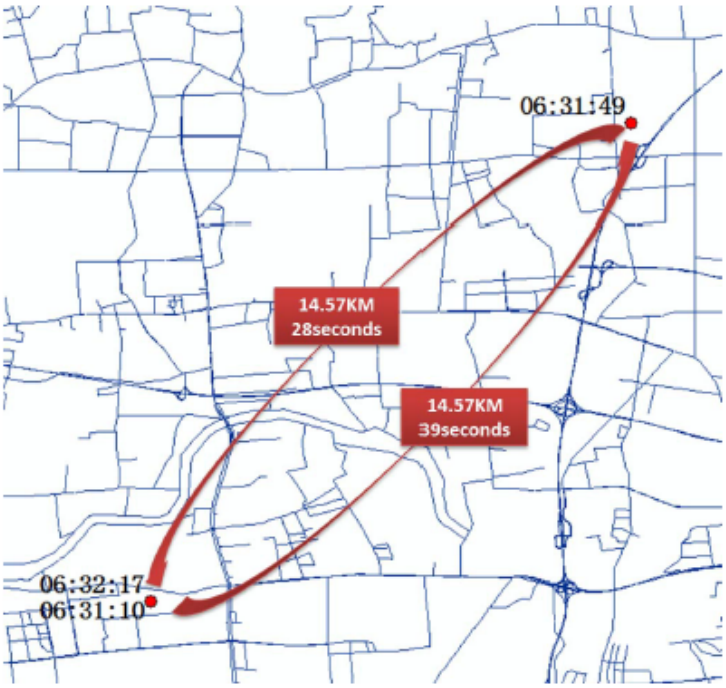


Figure 3.1: An example of reselecting a remote antenna, [21].

As previous authors mentioned, to this end several solutions are found in literature. A speed filtering algorithm, which takes every two consecutive event records with different locations and removes those whose speed of transition was above a certain speed is a common solution found throughout literature and well regarded in terms of its precision and recall. As an example, the work of [22] utilized a single threshold of 200km/h for all events. Another conference paper, [21] divided the cell tower reselection in two categories, remote cell tower reselections and frequent nearby cell tower reselection. For the first scenario, they find records where the cell tower is selected for only a brief period before the connection returns to the previous one, using a given distance and time threshold. The second one is more complicated and for that, they choose to use a clustering algorithm to group nearby towers and use the centroid as the new location for all those records. [20] implemented a similar solution to [22], however, the speed threshold was defined starting with a baseline of 120km/h and adapting it based on the time of day and region of the specific study area. In this thesis we will take an approach similar to [22] base implementation, as its simple yet effective solution proves to be enough to detect and mitigate these records in most cases.

3.2 Stop Points and Significant Places

Intuitively, we know that mobility involves going to and from a set of places, some of which are recurringly important to us and some of which are visited less often or only very rarely. A fundamental result, drawn from the analysis of CDR data [1], considers the probability distribution of the number of times locations are visited by users. In particular, such distribution is heavy-tail, with a few locations accounting for more than 50% of all visits by one user, a limited set of locations visited occasionally, while the long tail accounts for a large number of locations visited rarely or even once. Being able to discern those significant places in people's lives is an important aspect of characterizing human mobility.

The objective of this class of techniques is to extract from a given trajectory the sequence of locations that are visited by an individual, stops or stay points. In literature, the trajectories of concern are commonly only of spatial type, thus consisting of coordinated points. Solutions can be divided into two main groups, attribute-centric and pattern-centric segmentation techniques, as defined by [23]. Attribute-centric are the methods partitioning a spatial trajectory into a minimum number of segments in such a way that the movement inside each segment is nearly uniform with respect to some condition on movement attributes [23]. Attributes like speed, heading or curvature are all valid but the most commonly found attribute to segment CDRs is simply speed. For example, a stop can be defined as a segment along which the speed does not exceed a threshold value and whose temporal extent is lower bounded. A stay area would have a similar speed attribute condition but with a bigger temporal extent. The second group of pattern-based segmentation typically utilizes clustering methods for partitioning a trajectory in segments. Clustering methods, as explained in section 2.1 are either based on simple hierarchy, density or partitioning. Clustering-centric techniques are also used for the construction of semantic trajectories.

Using GPS data to search for these places is substantially easier as the sampling rate is much higher and the location accuracy is vastly improved. Therefore, this is by a large margin the best approximation of real movement. An advantage of such high sample rate is that it facilitates the use of interpolation functions to compute the likely position of the moving object for any instant between two consecutive sampled positions. The computed positions complement the captured positions, thus reconstructing continuity of movement. This can be seen in the work of [11] Semantic Trajectories Modeling and Analysis. Another advantage is that detecting changes in speed and direction is possible even at a small scale of movement and aggregating individual movement tracks into edges between nodes (stop points) of a flow network becomes trivial [24].

Diverting our attention back to Call Detail Records, certain studies make use of the attribute-centric approach, utilizing the coordinates and the timestamps contained in CDR records to infer stop periods and collecting their location. For example, in [21] a formula was devised to identify stop and move states, seeking periods of time with no variations of latitude and longitude. Both coordinate's gradients as well as the distance between consecutive events must be less than the defined thresholds for a stop to be identified. Significant places like home or work could then be identified by getting the stop location with the biggest time delta during the expected hours for a majority of the population to be at home or work respectively. Change point detection techniques [25] identify significant variations in trajectory patterns such as speed drops, significant altitude variations, U-turn, etc., which may bring additional values when searching for a change in mobility modalities. The main problem with utilizing such an approach with CDRs is that, as previously discussed, this information type is associated with spatial uncertainty, as cell towers have expansive and vastly different radius. Even if the records indicate no position change there is the possibility that the user has moved several hundred meters or more without changing the tower to which he is connected. So even speed or direction changes would only be detected at a macro scale, as small movements are, most of the time, not recorded.

Generically, it is found that the majority of application using CRDs prefer pattern-centric segmentation methods, because they possess another benefit. As discussed in the previous section, in CDRs, nearby cell towers serve alternately depending on a number of different factors including the user's location, signal strength and call traffic. This effect may generate many small stopping points that in reality belong to the same stay areas. For this reason, its common practice to aggregate the location points that are within a certain radius of distance to form location areas. As such, clustering algorithms serve the purpose of consolidating points that may represent the same location and avoid making a misinterpretation of the data. Moreover, some studies require the use of stay areas, a place where a user spent "a significant amount of time". It is different from a stay point, where no move is detected for a specific duration. Standard stay area detection methods are generally based on an iterative process aggregating intermediate stay points using geospatial bounds. Among existing clustering methods we can notice a prominence of density-based clustering ([21],[20],[23]).

One of the first attempts, found in [26], presented a clustering and regression algorithm combination to identify important places in people's lives. First, cell towers are spatially clustered using Hartigan's leader algorithm, a modified K-means clustering algorithm. Then home and work locations are inferred from cellular network data using a supervised method which achieves an accuracy of 88% within 3 miles of ground truth. However, the performance of supervised methods is strongly limited by the size of training data. In that study only 18 volunteers' home and work locations were used. For large-scale applications, user locations are usually anonymised thus unsupervised methods are needed to infer locations.

A study by [16] used a time-based clustering algorithm to pre-process the trajectory data, defined the processed location points as the stay points, and then used a grid-based clustering algorithm to extract important locations and defined them as the stay regions. The grid/based clustering has an advantage of a faster execution time at the cost of some accuracy, since the clusters are only made within each cell of the grid. [21] pre-processed the trajectory data by setting the movement thresholds to filter noise points, then used the DBSCAN algorithm to cluster the remaining location points and output the centroid of the clusters as important locations.This example of clustering loaction point with DBSCAN can be visualized in Figure 3.2. DBSCAN is still to this day considered as a very competent algorithm, capable of finding stay areas. Its recurrent appearance thorough literature supported our choice to use it in the methodology of this work.

Figure 3.2: An example of spatial antennas clustering with DBSCAN, [21].

An important work with some similar goals to our initial phase of development was found in [27]. They tackle the fact that people can be classified in two types: returners, those who are very regular in their daily mobility; and explorers, those who are inclined to break out of their daily mobility routine and explore new places. The latter type of behaviour remained relatively unexplored, and had very little in terms of literature so that was the proposed focus of their work, characterizing the places a person visits outside of normal routine. To achieve this they grouped the places visited by each individual by applying an unsupervised procedure, in conjunction with a relevance metric to the mobility traces extracted from CDRs. A K-means clustering algorithm with k = 3 served as the unsupervised method for aggregating the visited places into 3 distinct types based on the metric: (i) Mostly Visited Places (MVP), locations most frequently visited by the user; (ii) Occasionally Visited Places (OVP), locations of interest for the user, but visited just occasionally; (iii) Exceptionally Visited Places (EVP): non-routine places. In this thesis, due to the proximity of this research to some of our initial goals we replicated some parts of their approach. However, our works differ in the fact that they focus their attention on EVPs and we are more interested in the identification and classification of the MVPs and OVPs.

At this moment, it is important to point to the fact that positioning mobile users at a certain position with a high confidence level still constitutes a big challenge with CDRs. Approximating the user location in mobile networks has been investigated before. In the article [28], they proposed an algorithm for estimating the exact location of mobile devices in the cellular network. They use the Cell ID numbers to identify the base transceivers sectors. Each base transceiver sector covers an area; however, they do not specify the shape or the format of the area data they obtained. The team behind this research developed an app to simultaneously gather GPS and CDR information from volunteers. With GPS information they derived formulas for optimization of the coverage area location and an enhanced Kalman Filter with associated mobility models for estimating the exact location of the mobile users. Results produced an average error of 0.4 kilometers in the stay areas, a substantial improvement over previous works. In our study we also identify coverage areas by using the ID of corresponding sector cells, taking steps in this field of approximating the inferred user location to the truth.

3.3 Activity Identification (Points of Interest)

The majority of works described until know in this state-of-the-art chapter modelled spatial and temporal patterns of human mobility as a stochastic process around fixed points. The main shortcoming of these mobility models is that they overlook the activity (often referred to as the semantics of trajectory [11] a person engages in at a location within a certain time, i.e., they are not capable of explaining people's motivation behind mobility. On that account the objective of this class of methods presented in this section is to extract the semantic meaning of a location. In other words, to match a user location in the mobility traces with a motivation or a corresponding most likely activity. Something very present in many of these methods are Points of Interest. A Point of Interest or POI, at least in its basic definition as implemented by standard POI databases, is a place (comprising a location) associated with some semantic information (e.g., name, type, price, opening hours, address).

This process of semantic enrichment and disambiguation of places has been mostly seen with the use of GPS trajectories. Relatively recent studies, like [29], center their effort on the POI check-in issue, where the goal is to match the visit of a user to a specific registered location (e.g., restaurant, hotel, etc.). They propose a method to infer which registered locations were visited by users given their GPS trajectories and surrounding venue scores. The major issue of ground truth data was tackled by using a dataset from a previous work which contained GPS traces for 372 users with manually annotated Foursquare venue check-ins. The second problem was designing a probability inference model that predicts the check-in venue given a stop location and the nearby POIs. For this they adapt a general choice model describing the relative merit of a venue with respect to alternative candidates, taking as features the distance to the stop point and the number of other venues within a 500m radius of the stop point. The results show that there is room for improvement as their model was not very accurate, obtaining a low mean average precision of about 18% when comparing the results with the annotated data.

Trying to improve on previous researches, still with GPS, the authors in [19] proposed to perform user profiling by detecting significant places and their semantic meaning with external sources of information. Mapped information from four different geographical databases (HERE, Foursquare, Grand-Lyon, IGN) and nine different social and cultural databases (PreditHQ, International Showtime, Evenbrite, Songkick, Allevents.in, Meetup, Sportradar, 10times). The main implementation searched for overlaps between stay areas, calculated from GPS traces and POIs. That overlap was possible thanks to the geoshapes of each POI in the data. If no geoshape was available from the POIs database, then centroid comparison between the stay area and potential visit locations was executed. They used primarily the Haversine Distance (HD) between centroids, complemented by vertical distance, if available. For evaluation, a data collection campaign had users annotating their GPS traces with visited POIs. Their results show that on average the mean distance between a registered POI and a detected stay area was 21.2 meters but this value could go up to 80 meters.

These works perfectly showcase the main difficulty with trying to disambiguate stop locations without user input. Using CDRs further augments the issues, as the mobility traces are much less precise. A solution that seems to have been employed by some authors is classifying geographical areas by categories and obtaining the semantic meaning from the areas the user visits, instead of trying to match visit locations with a specific POI.

In [27] the metropolitan area of Milan was divided into regions by aggregating groups of close cell towers. In sequence, POIs obtained from Foursquare were used to classify these areas by the most frequent type of POIs present. In particular they achieved Regions of Interest, study area subdivisions classified by the top level categories of in the used POI dataset (e.g., Shop, Food, Nightlife, etc.). Results and conclusions were based only on the exploratory analysis of EVPs, finding for example the areas and categories most related to exploration and attempting to validate with common knowledge of exploratory activities. This work had influence in some of our initial methodology including tests with Foursquare and the use of top POI categories for classification of regions.

To further exemplify, the authors in [30], constructed a virtual grid reference, dividing the map in fixed cells of 500 by 500 meters. Then each grid cell is classified according to four main categories: Eating, Shopping, Entertainment and Recreational. To obtain that classification, the number of POIs associated with each activity category is recorded for each cell, creating an activity distribution map. Each cell activity proportion is then normalized to a value between 0 and 1. K-means with $k = 4$ (k = number of categories) is applied over these normalized values for all cells in the map to create 4 distinct groups. The interest here is to find the most probable activity category for each of the k clusters, for that a probability function is used, reaching the cell classification that can be seen in Figure 3.3. The next step was matching the grid cells with the user's CDR locations. For each user, daily activity patterns are collected over the course of the data collection period. Note that, in this study, they consider only weekdays (Monday, Tuesday, Wednesday, Thursday, Friday) as they speculate is that weekday pattern is different from weekend pattern due to typical work schedule and hence different daily activity sequences occur. The records are divided into temporal windows and to each temporal window is assigned the most frequent activity pattern for that time window in the whole period of data collection. No ground truth data and no validation were carried out as the authors fixate in their conclusion the difficulty and privacy concerns in obtaining this validation data.

This work provides a good basis to improve on. However, the chosen approach of using clustering on the grid cells based on their activity proportions means that this approach is difficult to replicate in a study like ours. Our study area is not well defined and we intend to create a solution that works whether the analysis is made in a small city or an big city. In this case results would be vastly different if clustering is applied in areas of different scales. Some other possible improvements could be taking into account the opening hours of each type of activity, using some kind of popularity or check-in count in the formulas for grid activity classification and not using a fixed cell size.

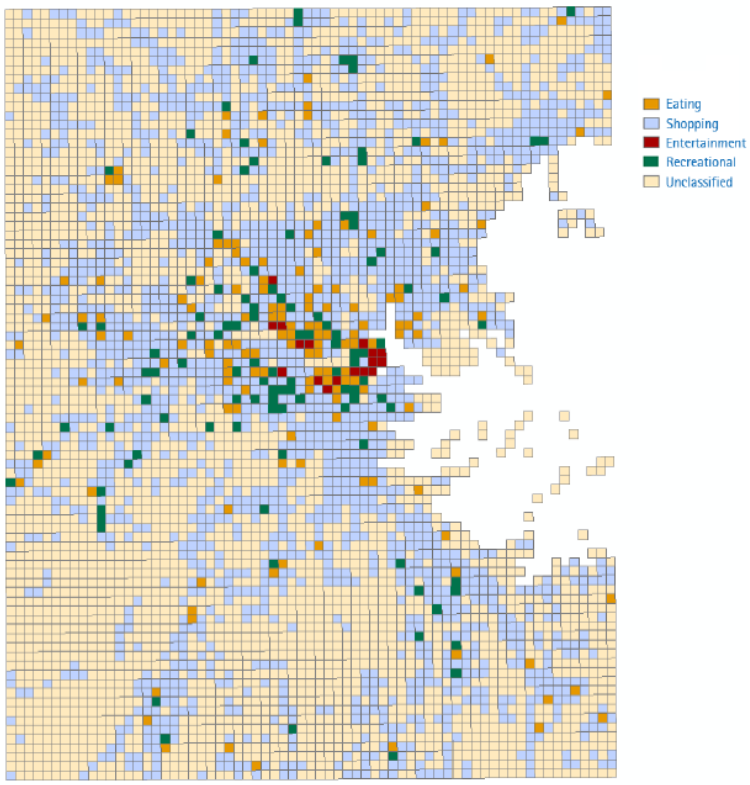


Figure 3.3: Resulting cell activity distribution map [30].

Similarly, but with the objective of matching traveler's internet searches with types of visited areas, in [31], a virtual grid reference was constructed to divide the city and each cell was to be characterized. In this case they test several sizes for the grid cells square sides (300m, 400m, 500m, ..., 800m) and use 10 different classification categories. For a cell k, the number of POIs in a given category is calculated, then this number of POIs of each category j is ranked over all cells, and the percentile rank is calculated as the percentages of cells that have lower number of POIs of category j than cell k has. As a result, each cell can be portrayed as a vector of the percentile ranks of all the POI categories. A method that can be done thanks to the small and closed boundaries of the study area, however would be harder to replicate in scalable application. Additionally, in this study, a hierarchical clustering algorithm is applied to the vectors of all cells. In a search for the optimum value, the number of clusters and size of grid cell are drawn in a graph for different values, using the Dunn index as optimization parameter. They conclude that, for their case, 500 meters is in fact the ideal value for grid cells and 6 clusters condense well the activity classification. Again, in this case they did not have access to validation data and they validate their results areas classification with personal knowledge from the study area, selecting specific examples of cells with clear primary activities.

Xu et al. in [32] focus on exploring the relationship between telecommunications and co-location, suggesting that mobile phone records help to predict the patterns of people who travel frequently to meet others. The goal was to quantify the temporal signature of geographical space in making friends, suggesting that urban spaces impact the spatial capability of connecting people. The relevant part in their research is when the study area, in this case the city of Singapore, was divided into 1423 clusters of mobile phone towers, or what they denominate as their grid cells. These were further grouped into 6 clusters using multi-level hierarchical clustering and a proposed metric for measuring the bonding capability of places. To obtain a better idea of the nature of places where friends are brought together, an analysis of the semantics of these clusters was conducted. To achieve this, they introduce five types of POIs, which are relevant to people's daily social interactions, from a dataset of the Singapore Land Authority: public commercial buildings, education institutes, shopping malls, sports centers and community. This allowed them to analyze the cumulative share of POI by each cluster and reach conclusions regarding why certain clusters had higher capability of bringing people together at certain time intervals.

In cases that geographical regions classification is the main goal, Points of Interest and CDRs are also used in conjunction, like [33], attempting to identify the activity centers for different activity types (such as working, residential and recreation). The geographical area was subdivided into Voronoi service areas of mobile phone towers and from them into 500 by 500 meters grid with the explanation that 500m is coarse enough to reduce the noise and detailed enough not to mix different areas. The number of people within each grid, obtained from CDRs analysis, was used to represent the activity density. A POI dataset allowed to find the number of venues within the grid, to proxy the POI density of the grid, as well as obtaining the number of categories is the cell, to proxy the functional integrity of the grid. Using the aforementioned values, they determine thresholds for classification rules with sample activity centers they had prior knowledge of. In the end, new activity centers are identified and validation was made through existing knowledge of the study area and comparing results obtained to that of the known reality.

A slightly different path is seen in the work of Yuan et al.,[34], where they primarily use the statistics of CDRs to classify geographical areas and POIs as a complement. Five geographical area subdivisions diagrams were taken in consideration and compared in this study: Raster layer, Voronoi Diagram, Road segmentation, Transportation Analysis Zone (TAZ) and Administrative. This departure from fixed cell approaches found the best results using the TAZ diagram. They classified each subdivision according to six main categories: Residential, Commercial, Park and Scenery, Office, Education and Mixed. The identification procedure is illustrated bellow in Figure 3.4.

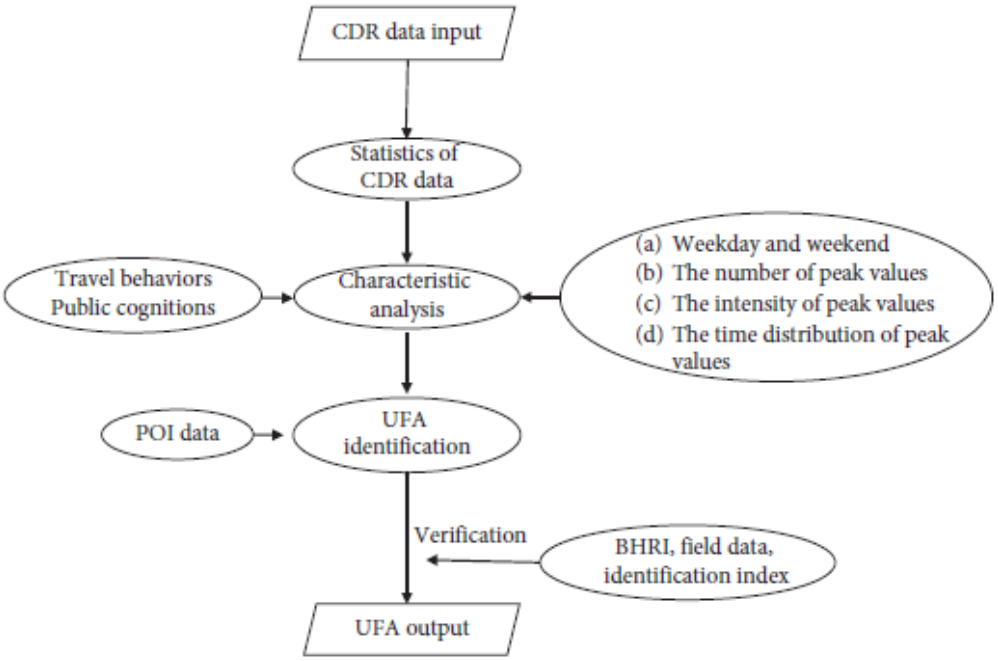


Figure 3.4: Methodology followed for geographical areas classification by [34].

First, based on CDR data, they calculate the parameters required for the index calculations. Second, the characteristics of the CDR clustering results are analyzed, including the weekday and weekend features, number of peak values, intensity of peak values, and distribution of peak values; this allows the travel behaviors and public cognition to be understood. Third, POI category density was used to complement the analysis of CDR events in each geographical division and based on that the identification results are modified. Validation was made through existing knowledge of the study area and comparing results obtained to that of the known reality. This interesting idea of using the distribution and peaks of CDRs to infer the classification of areas could be explored in more detail and if they shared their findings in detail could potentially be used to validate some future results.

Lastly one of the substantial more complicated models encountered was elaborated by [35]. Besides using the common 500 by 500 meters fixed grid, they combined mobile phone traces with an activity detection model not only allow them to illustrate the temporal dynamics in the Boston metropolitan area, but also to visualize the evolution of the spatial distributions of activities within the city. The complex activity probability equation took in temporal attributes (time of day and day of the week) as well as specific user features such as aggregate socioeconomic characteristics at the identified home grid cell. They also used two aggregate variables in the model to control for the weather effect: average precipitation and temperature in the Boston metropolitan area over the day. We took some inspiration from this paper's complex formula and inclusively adopted their time segments division. One day is divided into eight time segments to capture the intraday variations in activity participation: early morning (3–6 am), morning-peak hour (6–9 am), morning-work (9 am–12 pm), noon (12–2 pm), afternoon-work(2–5 pm), afternoon-peak hour (5–8 pm), night (8 pm–12 am), and midnight (12–3 am).

3.4 Evaluation and Validation of Results

As researchers turn to CDRs for studies on human mobility they find a specific difficulty in the lack of available large-scale annotated datasets to test their approaches and compare results with previous ones. Because this kind of data is associated with the previously discussed privacy concerns, they are not commonly available. Telecommunication service providers cannot publish their collected data publicly and as they are the only source, meaningful amounts of annotated data could only originate from them. This culminates in a lack of validation for specific works and forces some authors to recur to less-than-ideal methods.

A common method used by authors trying to identify home and work locations with CDRs ([21],[36]) is to compare obtained results with data from past population census or surveys. Dividing the study area into sectors and comparing the density of their model results with the existing ground-truth maps created from census. Although available, this ground-truth data is generally outdated. For example, in Portugal population census are only carried out every ten years. Even if not completely outdated, this data only serves a purpose for these focused works that intend to identify home and work and most of the time provides poor validation.

Another validation approach for home and workplaces that used collected ground-truth data was proposed by [37]. For each of the 11 volunteers, they collect real home locations to validate and the evaluate their methods with diferent parameters. From the places that their method identified as being the home location they selected the place that is closest to the ground-truth measuring the distance between both points. If the distance between the selected place and the ground-truth was below a certain threshold T , they considered to have found the place correctly and add 1 true positive score to the result. If there are no places, or if all the places are farther away than T from the ground truth, they considered it a false negative result. All the identified places that are not the true positive are false positives. In addition, to take into account the fact that distances in the city center are more "significant" than in the suburbs (in the city centers cell's radius are smaller, so the algorithm is expected to identify the place more precisely), the threshold T is adjusted to the average radius of cells in the area. However, their T was calculated using the formula $T = k + averageradius$. Where k was the radius of the home cluster. Results obtained with parameter $k = 1500m$ (they do not disclose $averageradius$) in home place recognition were: recall of 91%, 90% precision with 553m average spatial error and 1405m maximum spatial error. Although they only had access to 11 ground-truth location the applied methodology with the definition of true positives and false negatives was influential in our own validation. We additionally include the distinction between urban centers and non-urban areas as seen in this work to separate our validation results in two.

When it comes to stay areas identification, Hoteit et al. [14] propose an original technique to subsample GPS data in time, so as to mimic sparse CDRs, having a sample of real-world large-scale data from a mobile network operator as a basis. Their GPS data featured regular high-frequency position sampling, and covers the movements of 84 users worldwide for more than 18 months. With a lack of validation data to test their models this was their approach to simultaneously obtain ground truth and a sparse location dataset. Another research here previously referenced [28] faced with the difficult task of obtaining validation data, refocused their efforts to develop an application that recorded and sent the GPS traces and Call Records of volunteers to the research team.

In semantic disambiguation related works, the ones that use GPS data either performed several data collection campaigns, monitoring users around cities like [19], or had access to a dataset of annotated POI check-ins from previous works, like [29]. In [19] for example, results consisted of calculating the mean distance between a ground-truth POI and a detected stay area. Average was 21.2 meters but this value could go up to 80 meters. Most of those that proposed to do semantic disambiguation using CDRs, and were referenced in the previous section, did not contain a validation discussion highlighting the difficulties of obtaining ground-truth for that purpose.

When it comes to area classification, we can see that researchers chose a study area of which they possess personal knowledge ([31],[33],[34]). Since land use datasets do not generally contain the same classification categories that the authors intend, these are not regarded as a exact means of validation. As such, in the discussion section of their work, they end up comparing their study's results with land use data (if available), and their personal knowledge of the area's main uses and activities. Without proper validation data for this section of our work we were inspired by these previous examples to conduct a similar discussion on our results presenting examples of area classifications.

This page intentionally left blank.

Chapter 4

Data Description

This chapter presents, discusses and analyses the data obtained or gathered in this thesis. For each data type, one or various data samples are presented with the corresponding explanations. Exploratory data analysis, using statistical graphics and other data visualization methods, is conducted to summarize datasets main characteristics and verifying some assumptions.

4.1 Datasets

This chapter describes the datasets used in this thesis. Specifically, we used three main datasets. The first one is the complete dataset of Call Detail Records (CDRs), where events made from mobile devices are represented as entries in a table, identifying both the user and tower of connection. The second dataset is obtained to connect the CDRs to a point in space, the coordinates of the cellular towers. Both the CDRs and tower's reference file originate from the same data source, the telecommunication service provider. The final dataset is a tables of Points of Interest (POIs). Two sources were compared and tested for the purpose of obtaining this data, Foursquare and Facebook Places.

4.1.1 Call Detail Records

As previously stated, the study uses datasets of Call Detail Records from citizens of Portugal. Data was provided for this research by one of the largest telecommunications service providers in Portugal and before being made available the dataset was pseudonymized by the company. This means phone numbers were encrypted with a hash function, avoiding us from connecting our acquired knowledge to any single person, preserving the anonymity of user identity.

Table 4.1: CDR file sample.

dtime	cellid	dumpts	MSISDN	COD_UNICO
2020-08-31 19:07:46	609549	202009010400	0003F542C...	5C6029046...
2020-08-31 19:07:46	609549	202009010600	0003F542C...	5C6029046...
2020-08-31 19:07:46	609549	202009010000	0003F542C...	5C6029046...
2020-08-31 19:07:46	609549	202009010200	0003F542C...	5C6029046...
2020-09-01 10:11:28	609557	202009011000	0003F542C...	5C6029046...
...	...	...	...	...

For the development work carried out in this thesis a relatively small subset of data, comprised of nearly 36 thousand SIMs from the region of Coimbra, Portugal, was used. It corresponded to a period of 2 months, from 1st of September 2020 to 30th of October 2020, totaling 21 100 674 unique events. This amount of data was large enough to experiment, apply several state-of-the-art techniques and obtain representative results while still being acceptable in terms of run-time. A summary of this dataset is present in figure 4.1 and a sample can be seen in table 4.1.

Table entries did not consist solely of voice calls, a new type of event, named by the service provider as Snapshot, has been made available and included in the data. Snapshots consist of a way to fill the temporal gaps in CDRs, trying to mitigate the challenge of temporal sparsity. In fixed intervals of two hours, starting at midnight every day, the data provider added an entry replicating the user's last know position. This position is for all purposes a copy of the last event made, with the addition of a time column describing the hour value when the new entry was added. Although these new events do not give new information on user's position changes they provide an estimation in regular time intervals, just like the state-of-the-art methods of [14] and [15].

The entire records table is constituted by two user identification fields ($MSISDN$ and COD_UNICO), the timestamp of the last event beginning (*dtime*), the last time of snapshot communication (*dumpts*) and a unique id for the cellular tower antenna (*cellid*).

Additionally, we had access to two reference tables. The first one provided the connection between antenna identification numbers, *cellid*, and the corresponding cell tower location coordinates. This table is discussed in the section that immediately follows this one. The second table was a reference table for home address, containing ground-truth data for a sample of the study population. This table is later described in section 6.2.1.

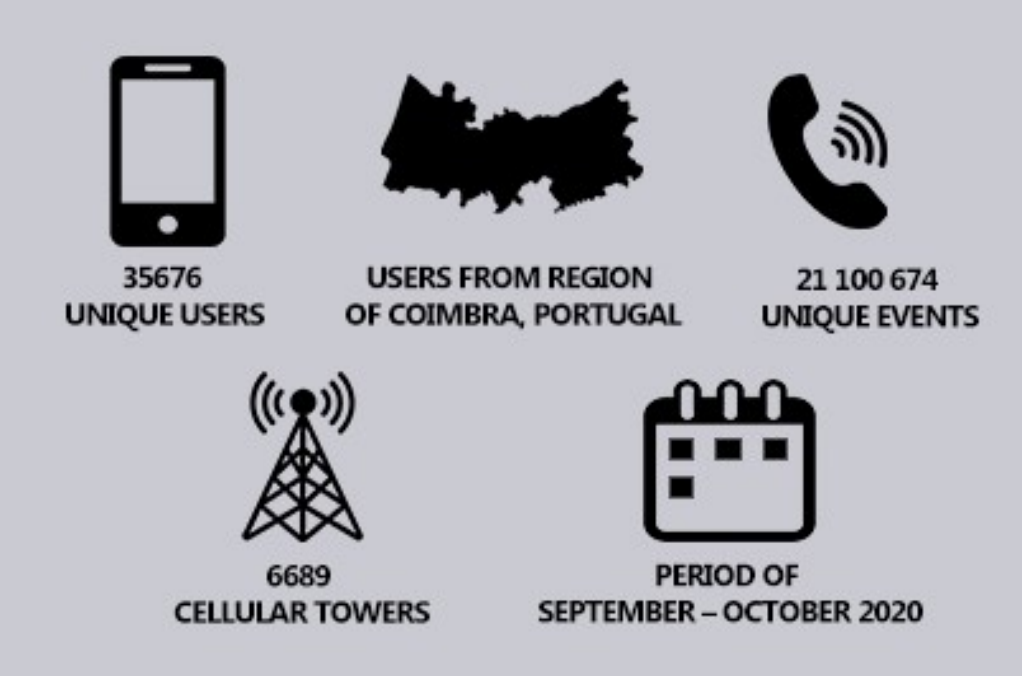


Figure 4.1: Key statistics of the dataset.

4.1.2 Cell Towers Reference

The second dataset, as important to the rest of the work as CDRs, is the cellular towers reference tables. In order to utilize the Call Detail Records supplied by the data provider, which (described is section 4.1.1) do not possess information about geographical coordinates, a secondary auxiliary dataset was necessary. This dataset came to facilitate the matching between the cells identification number and an actual real world position. As might be expected, this data was also provided by the telecommunication service provider.

This set of information was subjected to several versions throughout the period of this study, each one increasing in the amount of information fields present. These changes where sometimes requested, as a means to increase our understanding of the cellular network and explore different solutions and optimizations. Other times they where added as we converged on a path of development that needed specific data.

The first and simplest version of this dataset contained the following fields exemplified in table 4.2, including the identification number of the antenna (or COD_CELULA), and the Cartesian coordinates of the tower in which the cell is located.

This information is the foundation that was the needed for the initial phase of development and was used mostly during the first semester.

The following version was extensively more detailed, containing a description of the antenna's area of service (*nome_celula*), as well as fields of Portuguese territory subdivisions such as *Freguesia*, *Concelho* and *Distrito*. By having this information present it was now possible to filter data with great efficiency and to focus our study on specific areas.

Table 4.2: Cell Reference table Version 1.

COD_CELULA	VAL_LATITUDE	VAL_LONGITUDE
16886	38.648818000	-9.037172000
16891	38.657520000	-9.062470000
5357	38.648880000	-9.044730000
9941	38.656310000	-9.068270000
...	...	...

It was at this point that a possibility for obtaining a more detailed position of the user's location was discussed, thanks to our fortnight meetings with the client and presentation of ongoing work. New information fields were needed that contained for each antenna the initial and final angle of action (*angulo_ini_celula* and *angulo_fim_celula*), in degrees, as well as the estimated range value (*raio_cobertura_antena*), in meters.

The last modification to the reference table added an information column with the corresponding antenna service type (e.g. 2G, 3G or 4G). This was appended to the remaining table as we wanted to run tests excluding certain service types and study if it produced a positive change in the results.

A sample containing the final version (Version 4), with all the new columns can be seen in table 4.3.

4.1.3 Points of Interest

A Point of Interest (POI) is an entity of interest with well-defined location. Points of Interest can range from famous landmarks (e.g. museums, churches, towers), natural attractions (e.g. bays, coasts, waterfalls) to commonplace spots (e.g. coffee shops, taverns)[38].

In order to semantically enrich the collected CDR data we intend to infer a person's activity given her/his location traces. To do that we need to create geographical area subdivisions, associate these areas to a list of POIs they intersect and infer the most probable activity by mapping each POI to a specific activity. The activity we obtain from the category, an information field present in most datasets under the specific attributes of each POI.

Generally each Point of Interest has as main attributes the name, category and Cartesian coordinates. Some databases can include other relevant information such as top level category, opening hours and popularity features such as check-in counts or user ratings.

Table 4.3: Cell Reference table Version 4.

codbk_celula	nome_celula	latitude_wgs84	longitude_wgs84	freguesia_antena
7006	STO TIRSO CENTRO 2	41.343286	-8.478648	União das freguesias de Santo Tirso Couto (Santa Cristina e São Miguel) e Burgães
30440	COIMBRA TOVIM FDD 1	40.21462	-8.396	Santo António dos Olivais
31709	GRIJÓ OESTE FDD 1	41.032276	-8.602924	União das freguesias de Grijó e Sermonde
520747	PRIOR VELHO LD 3	38.788436	-9.124982	União das freguesias de Sacavém e Prior Velho
388363	RIO MOURO LA 1	38.78308	-9.342106	Algueirão-Mem Martins
10867	ÉVORA MISERICORDIA 2	38.5698666	-7.9071888	União das freguesias de Évora (São Mamede Sé São Pedro e Santo Antão)
36023	COVA PIEDADE FDD 3	38.673533	-9.160344	União das freguesias de Almada Cova da Piedade Pragal e Cacilhas
...	...	...	...	...

concelho_antena	distrito_antena	angulo_ini_celula	angulo_fim_celula	raio_cobertura_antena	tecnologia_ceLula
SANTO TIRSO	PORTO	115	235	1300	2G
COIMBRA	COIMBRA	57.5	237.5	8200	3G
VILA NOVA DE GAIA	PORTO	325	85	2100	3G
LOURES	LISBOA	165	285	927.685	4G
SINTRA	LISBOA	360	110	331.462	4G
ÉVORA	ÉVORA	160	280	1300	2G
ALMADA	SETÚBAL	200	117.5	700	3G
...	...	...	...	...	...

POIs can be collected from various online sources and are frequently available for free through Application Programming Interfaces (APIs) and online map service providers. Used POIs can be extracted via a single API or obtained and aggregated from several sources. In our particular case it made more sense to use a single source since places are classified into various categories covering a variety of subcategories, and there are overlapping problems in different datasets, so it would be necessary to reconstruct and reclassify the POI data to join two or more sources.

Foursquare API Data

Foursquare is a social networking service available through the web and a mobile application for most smartphones. It allows users to connect with friends by seeing each others visited places, has a well regarded places recommendation system and allows to search and read reviews of any registered place, among other features. It is more oriented towards tourists attraction and leisure POI categories (e.g. restaurants, bars, museums), lacking in less checked-into categories, like a doctor's office or grocery stores.

Taking inspiration from some of the works in literature and known studies, for example, Quadri et al.[27] and Qihang et al.[29], Foursquare's seemed a good first option for providing this data.

Their free access API allows to retrieve a specified number of points by area searches or provides detailed information about a single POI through a specific search by individual id number. The API, however, has the downside of limiting the amount of requests made in a day while maintaining a free access and returns results randomly in a way that makes it difficult to store an offline version. Although anyone can make an account for free the base functionalities in terms of number of accesses are quite limiting without choosing one of the available payment plans. The limits are 950 regular API requests per day and 50 premium API requests per day for free tier accounts, jumping up to 99 500 regular calls and 500 premium calls per day after a credit card is confirmed. Regular requests include basic venues location data, category, and ID. Premium requests include rich content such as ratings, hours, photos, tips, menus, etc. For example, search for a specific venue around a given location is regular, but learn more about a specific venue is premium.

Facebook Places API Data

Facebook is probably the most popular social networking site that makes it easy for people and or businesses to connect and share with family, friends and clients online. Facebook Places is an associated geolocation service built into Facebook that is designed to help users share their favorite spots and discover new ones. Users can "check in" at various locations, from cities to small stores. Additionally users are given the ability to create a new POI if the one they intend to 'check-in' or review does not already possess a Facebook Page. Business owners can claim and certify the pages created by a third party by following a verification process.

Table 4.4: Sample of the Facebook Places POI table.

name	check-ins	hours
Restaurante Aviz	425	{"key":"mon_1_open","value":"09:00"}{"key"...
AZULMIR	15	{"key":"mon_1_open","value":"09:00"}...
B-Culture	0	{"key":"mon_1_open","value":"09:00"}...
...	...	...

latitude	longitude	category	city
39.82468	-7.4915	Portuguese Restaurant	Castelo Branco
40.43211	-8.72678	Wholesale & Supply Store	Mira
41.45011	-8.33808	Medical & Health	Guimarães
...	...	...	...

The main benefit of this data source when compared to Foursquare is the bigger outreach of the Facebook platform and as such the amount of POIs is increased as expected. There is in fact a representation of categories that are not present in Foursquare's database, including organizations, societies, finance and healthcare. These categories, although not as important for tourism or leisure, are important to infer everyday mobility motivations for resident population.

Furthermore a dataset had already been constructed for the whole country in a previous work by AmILab [39], allows the access to an extensive offline database, an important factor for creating our scalable algorithms. Also, by observing both datasets, it was perceived that a bigger percentage of Facebook POIs contained information on opening and closing hours, something that we further enhanced as is described on the chapter 5. Unlike Foursquare, the column fields of the POIs (see table 4.4) do not contain the top level category in the categories hierarchy. This means that a separate dictionary file was needed that contained the category name, unique id and the parent category id.

Table 4.5: Dataset analysis on number of events per user.

User Count	35 676
Mean Events	591.453
Std	178.591
Min	1.000
25%	522.000
50%	664.000
75%	719.000
Max	1458.000

4.2 Exploratory Data Analysis

This section intends to use statistical graphics and other data visualization methods, to summarize the main characteristics of both CRDs and towers datasets. The primary goal being to maximize insight, make useful discoveries and verifying some assumptions.

4.2.1 User Records

Analyzing the obtained Call Detail Records in regards to events made by each user we can ascertain the values present in table 4.5. For the 35 676 individuals there is an average of approximately 591 unique events recorded with a standard deviation of 178. The number of users with more than 800 events is very small with the maximum recorded being 1458.

Figure 4.2 represents a histogram of the events per user but instead of a count in the Y axis a kernel density estimation is used. As can be seen from both the quartiles described in table 4.5 and figure 4.2, the events do not follow a normal distribution. In fact, having an appreciable positive skewedness while peaking at the 724 events mark. This is not exactly ideal since our interest is in individuals with a high event count in order to facilitate testing and development of an accurate pattern discovery methods. Our thought process being that the more active an individual is in terms of records the more accurate we can extract and analyse his/her mobility. However, users with a low event count can be used to validate the systems capabilities on all situations.

The period of September/October of 2020 represents, at the time of writing, the best possible chance for normal population movement patterns from the data that can be made available to us. Given that, in the case of Portugal, it is from a more relaxed Covid19 confinement period. Even though we know it will never give us an accurate indication of pre-pandemic mobility, however it is the closest we have gotten since the data collection for this project began.

In search of a better indication of mobility in the data we conduct an analysis of unique cell towers visited by each user. This would serve as a better indication if this data contains potential for mobility studies, or is compromised by the atypical situation lived throughout the year of 2020.

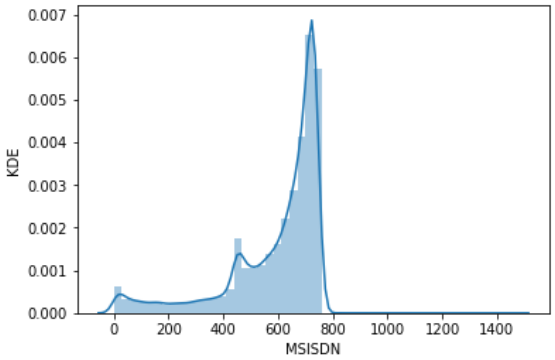


Figure 4.2: Histogram of number of events per user.

Table 4.6: Dataset analysis on unique cell towers per user.

User Count	35 674
Mean Unique Towers	25.309
Std	18.406
Min	1.000
25%	12.000
50%	24.000
75%	34.000
Max	224.000

The information present in table 4.6 and figure 4.3 shows that the average user has 25 different locations recorded and the standard deviation is relatively high which should be expected as there is a big spread of values. This should certify that although not an ideal timeframe for this type of study the dataset contains a good number of visits for each person.

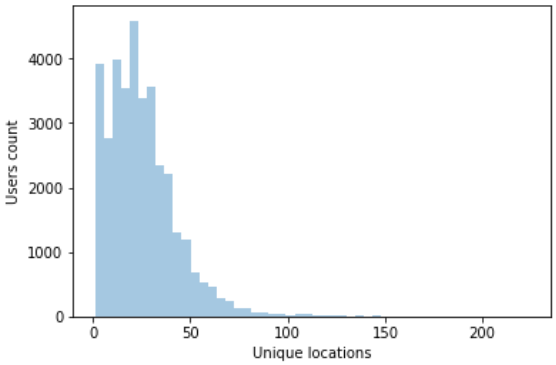


Figure 4.3: Histogram of unique cell towers per user.

4.2.2 Cellular Towers

Portugal has an estimated territory comprised of 92 212 km², including the autonomous regions of Madeira and Azores. Cellular towers, however, are not evenly distributed throughout that territory. There are clear differences between the tower density of some districts. This density is proportional to population densities. More populated regions, such as the cities of Lisbon and Porto, require large groupings of towers to support the amount of network usage. Coastal regions in general have increasingly growing population, which is reflected in the cellular network infrastructure present in these areas. Figure 4.4 presents both towers locations and a heatmap from those same locations in the continental region of Portugal.

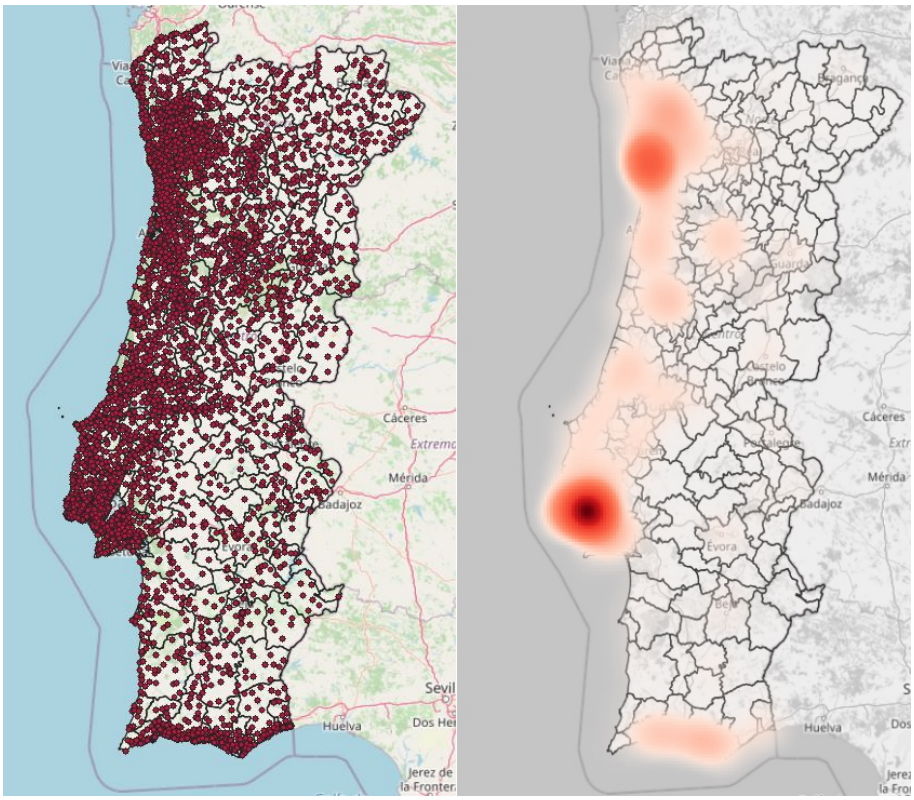


Figure 4.4: Cellular towers distribution and its respective heatmap for continental Portugal.

Despite having the references of all 6689 cell towers our study area does not consist of the whole country. We focus our work on the district of Coimbra, as it is more manageable in terms of the amount of obtained data for developing and testing. Although it does not possess one of the largest tower groupings, like Lisbon or Porto, it possesses some value in the fact that the author and the whole team at AmILab have intimate knowledge of the study area. Visualizing the towers present in the district (figure 4.5) we can distinguish the urban center of the city of Coimbra, positioned in the center, with a large number of points. This large cluster is surrounded by less dense towers locations.

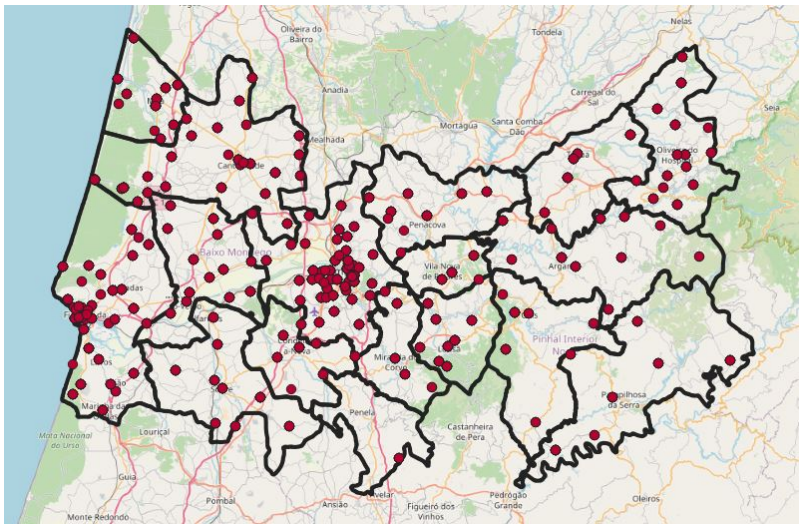


Figure 4.5: Cellular towers in the district of Coimbra.

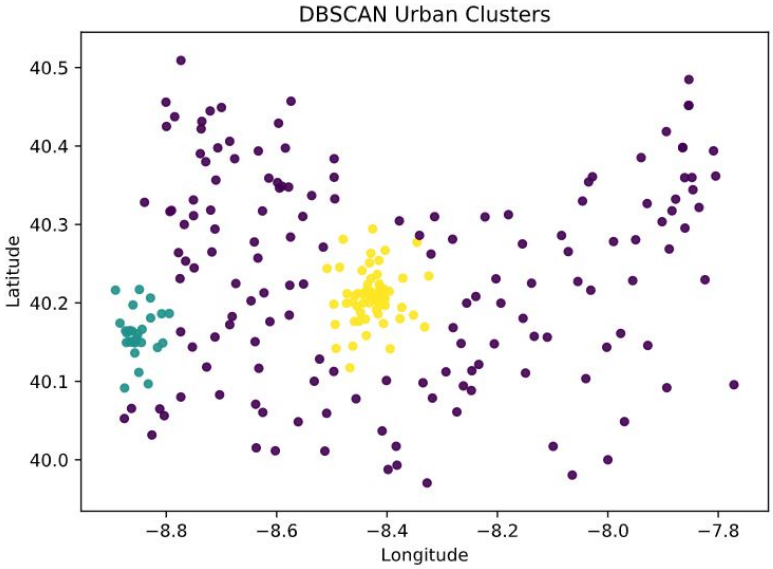


Figure 4.6: Detection of urban clusters using DBSCAN algorithm.

The distinction between dense urban centers and less dense areas is important in this thesis, as methods involving pattern identification using CDRs benefit from the higher density of cellular towers. These should be accounted for in evaluation of our methods. This analysis intended to group towers into urban clusters, making the contrast between urban and non-urban areas. For this we used a density-based clustering algorithm, more specifically DBSCAN (section 2.2.3), to split the antennas into clusters, representing high density urban areas and outliers representing less dense rural areas. Results are visualized in figure 4.6 with all points plotted according to their Cartesian coordinates and color coded by cluster. Most towers, in purple, were categorized as outliers, not belonging to any cluster. Two urban clusters were found for the cities of Coimbra and Figueira da Foz, identified by the yellow and blue colors. This analysis later has an important role in validation and evaluation, inspired by [37], we separated the results by urban and non-urban areas as tower density is a relevant feature.

Table 4.7: Dataset analysis of antenna's radius.

Antenna count	735
Mean radius	1976.79
Std	1952.77
Min	76.11
25%	785.05
50%	1300.00
75%	2462.77
Max	10500.00
Mean radius urban	2084.80
Mean radius rural	2961.29

Doing an analysis of antenna's signal radius in the district of Coimbra (figure 4.7) we can see the majority of radius are bellow or equal to the 4000 meters value. The histogram is more clearly described by the values in table 4.7. The mean radius of antenna's in the study area is 1976.79 meters, however the mean value for antennas in the urban clusters is 2084.80 and 2961.29 for the non-urban/rural.

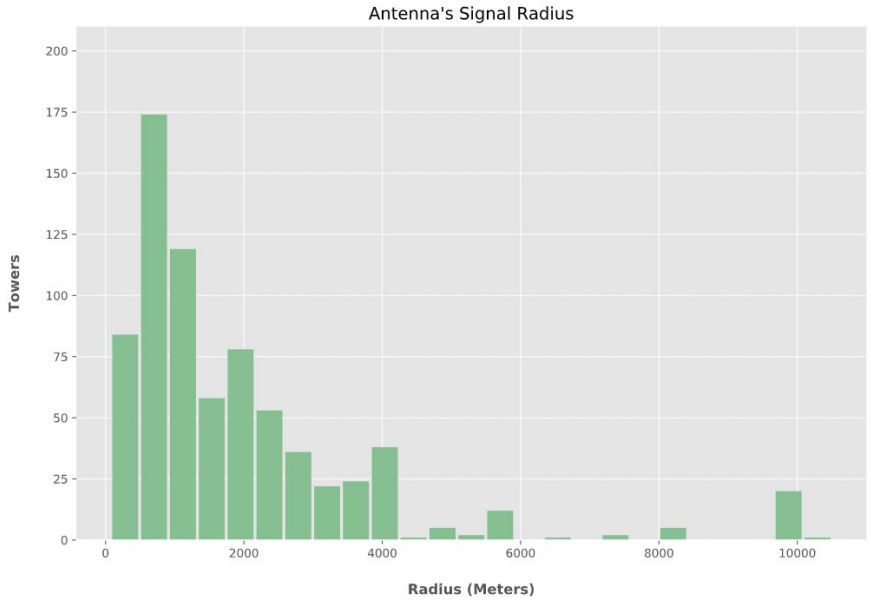


Figure 4.7: Histogram of antenna's signal radius.

Another analysis possible with the towers reference file would be the distribution of towers/records by service types (e.g. 2G, 3G, 4G). An initial hypothesis was that types of service introduced first, like 2G, could be less prevalent and associated with more location uncertainty, due to higher radius antennas, negatively influencing the results. We created graphs of the distribution of service types to check this assumption. Figure 4.8(a), is a pie chart of the distribution of service type by antenna's inside the study area. The overall percentages are well balanced with a slight tendency to increase as we move up to newer service types like 4G. Figure 4.8(b) is the pie chart representing the distribution for service types in the user events of the CDR dataset. Percentages are very similar to those of the antenna's with a minor loss for 4G in favor of 3G and 2G. By these graph alone we conclude that we cannot exclude any single service type as they all have a large prevalence in the datasets.

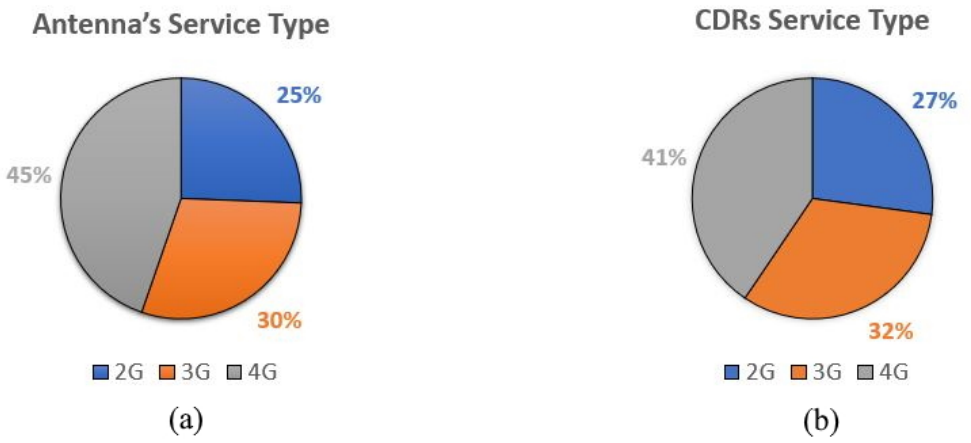


Figure 4.8: Analysis of service type.

This page intentionally left blank.

Chapter 5

Methodology

This chapter presents the discussion of the research methodology on work developed along both semesters. Every aspect of the research conducted is presented, including the reasoning and selection of methods regarding pre-processing techniques, clustering algorithms, activity classification and area selection.

Several previous works discussed in chapter 3, serve as the basis of inspiration for this work with the ultimate goal of building on the state-of-the-art methods. From the study and analysis of the state-of-the-art research we created an initial road map for experimentation with clear goals for the first and second semesters. The work can be divided into several sections with the final goal being to create a system where given a user's CDRs as input, a detailed table of each routine place and its classification is given as output.

The methodology here presented was regularly followed by the AmILab research group, through the weekly meetings. The telecommunication service provider, client and recipient of the final outcomes, also accompanied the project by way of status meetings, held every two weeks. This provided for constant feedback, leading to several iterations and adjustments of the planned methods to align common goals between the research work and the business needs of the client.

5.1 Call Detail Records

As explored in chapter 4, section 4.1.1, the base Call Detail Records and reference tables for the entirety of this project were provided to us by a major telecommunication service provider. Records are confined to a two month period, between September and October 2020, and data collection was made for users who had a majority of mobile events in the district of Coimbra, Portugal.

5.1.1 Data Exploration and Preparation

Initially, from 35 676 users a total of 21 100 674 records constituted the dataset. These were a mix between event-driven and network-driven, which means that not every single event was attributed to the user making a call or message.

Firstly, after opening and visualizing the tables, they need to be ordered by user id ($MSISDN$ field), counting the number of dataset entries for each. As can be seen by the data exploration in section 4.1.1 there were some cases where users had a very small number of events. Even users with less than one event per day. As expected these will add little to no information for our research purpose, given we want in fact a larger number of events in order to infer spatial patterns, only delaying and creating unnecessary overhead when running the chain of algorithms. As such, a simple function was created that, receiving an event threshold as input, filters out all users with a number of events bellow that threshold. For example, removing users with less than one event per day resulted in a reduction of 0.12%, or 25 858 unique entries.

The next step consisted of merging the user's CDRs table with the towers reference table. Since both tables have a common field (*cellid*), this was the matching key. After the operation we checked to find and remove null values in both the latitude and longitude columns, which could be the result of a missing or defective *cellid* in the towers reference file. In total 379 491 events were removed by missing latitude or longitude values, around 1.8% of the dataset.

Following the addition of the coordinate columns, we were inspired by works such as ([35],[40]), to retrieve additional fields (e.g., day of the week, workday/ weekend, etc.) from the existing information and add them to the table in new columns. An integer for day of the week (from 0 - Monday to 6 - Sunday) and a boolean value for workday or weekend (0 being the workday) were obtained from the timestamp columns using existing functions in python's 'datetime' library. This extra information will help to filter records in the next steps and finding patterns in the data with regards to user's places. Furthermore we adopted the time segments division found in [35]. So for each event, taking the timestamp, we verified to which time segment it corresponded and added that segment in a new column. One day is divided into eight time-segments to capture the intraday variations in activity participation: early morning(3–6 am), morning-peak hour (6–9 am), morning-work (9 am–12 pm), noon (12–2pm), afternoon-work(2–5 pm), afternoon-peak hour (5–8 pm), night (8 pm–12 am), and midnight (12–3 am) [35].

5.1.2 Load Sharing Detection

We also address the issue where cellular tower reselection happens in the middle of the call, or in very quick succession, due to automatic network load balancing. With this in mind, distances between the network towers involved in consecutive CDRs entries and the travelling speeds of the users were estimated. For the detection of the Load Sharing effect a speed-based method was implemented, a sequence is identified if the switching speed exceeds a given threshold. We set the value at 200km/h in conformity with the work of Iovan et al. [22]. This approach is described in the pseudo-code in algorithm 1. Results obtained are analyzed in chapter 6, section 6.1.1.

Algorithm 1: Load Sharing detection and removal.

Input: CDR table and speed threshold (s)
Output: Filtered CDR table

```
Filter the dataset with a query for the following condition: consecutive
  records, of the same user, on the same day, in different locations;
for each record pair do
    Extract the time values from the timestamp column;
    Calculate the time difference between the two records (t);
    Extract the coordinate values from the latitude and longitude columns
      and make two points;
    Calculate the geodesic distance between the two points (d) (Default
      WGS 84 ellipsoid);
    Calculate the speed of transition (st) using time (t) and distance (d);
    if speed of transition (st) is above threshold (s) then
        Flag the two records as Load Sharing and remove them from the
          data ;
    end
end
```

To calculate the distance between two pairs of Cartesian coordinates we use the geodesic distance between two points. A geodesic is the shortest path between two points on a curved surface, analogous to a straight line on a plane surface. This calculation is one existing function in the geopy library, with the input being two separate coordinate points. The default ellipsoid used for the calculations is the World Geodetic System 1984 (WGS 84) datum surface, widely used for mapping and satellite navigation and considered to be the most accurate.

5.1.3 Home and Workplace Detection

One of the main parts of this project is the accurate identification of each user's home and workplace locations. These will most likely be the places where people spend the majority of their time and, by consequence, could amount to a large portion of their mobile records. Finding these locations accurately is important because it allows us to focus our attention on relevant records for our research, as well as speeding up the forthcoming algorithms by diminishing the size of the dataset. Thankfully this topic has been a subject of many prior studies and there are proven methods with good accuracy.

Based on the works explored in chapter 3 it was decided to use a mixed approach of time filtering with a clustering algorithm [36]. Firstly, we select the temporal intervals of search when its most likely for someone to be found in the places we want to identify. In this case home time interval was defined as the interval from 7PM to 9AM as per [36]. However because they do not experiment with detection of the workplace location. In this case we defined workplace time as the interval from 9AM to 5PM, a common 8 hour period for normal day workers. Workplace time CDRs where additionally constrained to workdays, given the extra field that we created during pre-processing.

The best choice of clustering for our application, given the research conducted on the state-of-the-art methods, is a density based one, specifically we opted for Density Based Spatial Clustering or DBSCAN ([21],[20]). This algorithm does not require pre-specifying the desired number of clusters, which is important for our case as we want to create an indiscriminate amount based on distance between the points. It is also more efficient at handling outliers and the clusters can take any form, unlike more popular spatial algorithms like K-means clustering. As explained in section 2.2.3 the algorithm has two main inputs that can be optimized, the threshold distance, or Epsilon (Eps), and the minimum number of points (MinPts) needed inside that threshold to define a cluster. The clustering function was implemented using the DBSCAN from the sklearn python library and the initial values we used were Eps = 0.01 and MinPts = 2 [21]. However, when we obtained validation data for home location further optimization was possible to these parameters. As we are working with points in Cartesian Degrees, the Eps value is also in the same unit, with 0.01 roughly approximating to 787 meters in out latitude [41],[42].

The DBSCAN algorithm is applied in an individual manner to each user's records. This process is done twice, changing the temporal interval depending on which location we intend to detect, home or workplace. Given an array of coordinate points as input, the algorithm outputs an array of integer labels, each label corresponding to a cluster. Every location with the same label belongs to the same cluster, with the exception of the negative one label (-1), which corresponds to outliers. If more than one cluster is found, the cluster with the highest number of events is selected as the most likely location. Figure 5.1 allows for the visualization of the resulting location clusters for one user's home hours, with each cluster being identified by a unique color.

In addition to returning the cluster of locations that are identified as either home or workplace, the centroid of the cluster, is calculated and given in the output as well. The centroid being the arithmetic mean position of all points in the cluster. This is the location later compared with the ground truth data.

The *cellid* values that belong to the detected clusters are removed from the user's CDRs dataset. In this way, excluding these two important places, we stop them from being mislabeled later on or in any way negatively impact the results of subsequent methods. All steps for this approach are described in pseudo-code in algorithm 2.

Algorithm 2: Home and Work Detection Algorithm.

Input: User CDRs dataset

```
Filter the user's dataset for events only on the designated time interval and
  weekdays;
Create a new dataframe only with the coordinate columns;
Run the DBSCAN algorithm with input: Location's coordinates, Eps,
  MinPts;
if Only one cluster is found then
    Calculate the centroid of the cluster;
    Return the cluster and centroid;
else
    Count the number of events in each cluster;
    Choose the cluster with most events;
    Calculate the centroid of the cluster;
    Return the cluster and centroid;
end
```

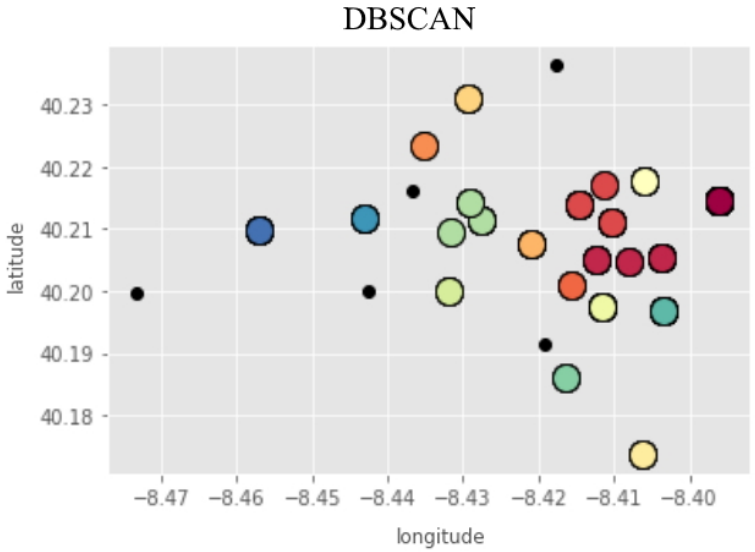


Figure 5.1: Clusters obtained with DBSCAN for one user's home hours.

5.1.4 Other Routine Locations

After identifying and excluding homes and workplaces from the individual user's dataset we are left with the other cellular towers the user was connected to in the 2 months of the data collection. Having these remaining location coordinates we then need to understand which ones possess more relevance to the daily routine, those that are more frequently visited and account for a substantial time expenditure.

As chosen to detect home and workplace, DBSCAN clustering could also be used to find other routine locations. Without the restrictions imposed with the home/workplace hours and by keeping all clusters, instead of singling out the cluster with most events, it would be a good candidate solution. The issue found with using this density based clustering on the CDRs coordinates is that we would lose additional precision in pinpointing the exact user position. Antenna locations already have a large uncertainty when it comes to matching the user position and clusters consisting of several antennas would increment the challenge substantially. This is acceptable when we intend to identify home and workplace, however, for routine places we want to retain the maximum precision possible as these are intended to be analysed further for activity classification.

Inspired by the work of Quadri et al., [27], that divided user's locations in classes of importance with respect to the number of unique visit days, a similar approach was experimented. The three classes are: Mostly Visited Places (MVP), locations most frequently visited by the user; Occasionally Visited Places (OVP), locations of interest for the user, but visited just occasionally; Exceptionally Visited Places (EVP): non-routine places. To distinguish the most and occasionally visited from the ones rarely visited we tried two main approaches.

The first approach consisted of estimating the total time spent in each locations. This phase was experimental as we planned to apply the relevance metric used in the original work, or possibly a combination of different ones to improve results. Since the dataset does not contain a field with duration of events, in order to know the total time a user spent at any given place, we calculated the time difference t_{ij} between consecutive events $iandj$. That time difference t_{ij} was then attributed to the position p where the events were recorded. The resulting times t_{ij} were then summed for each user's individual position p. Once the time metric column was calculated we used it as input in a K-means clustering algorithm. This was done as we intended to separate the locations in two groups, where the threshold for separation varies on a per user basis. To this end, an unsupervised learning algorithm such as K-means clustering is a good method. As explained in section 2.2.2 K-means requires a specified number of clusters beforehand, so in this case the input value was k=2, Most Visited Locations and others that are less or rarely visited.

We were aware that the initial approach of calculating the time spent at each location could lead to inconsistent results. Since this dataset of call records does not contain an event duration column there is some uncertainty associated with estimating this based on only one timestamp. With these challenges in mind, this methodology was eventually replaced.

The second approach saw the user points divided in three groups instead of two, comprising the most, occasional and rarely visited locations, like in the original work [27]. In addition, we utilised the original location relevance metric instead of the total time spent calculation. The relevance metric was calculated based on the number of unique days the location appears in the CDRs. So the relevance of a location l for a certain user u - $R(l, u)$, is calculated by the number of unique days the user visits the location $d_{visit}(l, u)$ over the user's total number of active days $d_{total}(u)$ (see equation 5.1).

$$R(l, u) = \frac{d_{visit}(l, u)}{d_{total}(u)} \tag{5.1}$$

Several factors can affect the activity someone takes part in a location, users can use the same location in distinct ways depending on the day and time. As our main goal is not only to detect routine locations but also to infer activities, the relevance metric was modified to accommodate this need. Before advancing to the metric calculation we grouped user CDRs by coordinates, time interval and type of day (workday/weekend). Instead of counting the unique days the user visited a set of geographical coordinates (location l), we counted the unique days the user visited l in time interval tm and type of day w. The final metric for calculating the relevance of a location depending on time and day type ,$R(l_{tm,w}, u)$, was the one presented in equation 5.2.

$$R(l_{tm,w}, u) = \frac{d_{visit}(l_{tm,w}, u)}{d_{total}(u)} \tag{5.2}$$

As with the previous method we used the newly calculated metric column as input to a K-means clustering algorithm, this time with input value k=3 to obtain the three distinct groups. Figure 5.2, a 3d scatter plot, shows coordinate points clustering by the relevance metric for one selected person in the data. Note that the Z axis represents the relevance metric while the X and Y are latitude and longitude respectively. Purple color coded point, with the highest relevance score would are MVPs, with orange points being OVPs and blue points EVPs.

5.2 Points of Interest

One of the main branches of this thesis is the classification of areas in terms of most probable activity. Once the user's CDRs were analysed, locations points were found that did not contain any meaning other than geographic places where the user is detected regularly at a certain time and type of day. Work in this section ahead had the objective of giving better insight into possible motivations behind the mobility by classifying these areas with the most likely activity.

To achieve this, we resorted to Points of Interest (POIs) as the base data. These points would not be obtained from the telecommunication operator, as they do not collect or store this information, instead it was necessary to use a specialized database. Tests were made with data from two different geolocation social networks, Foursquare and Facebook Places, with the pros and cons weighted for each.

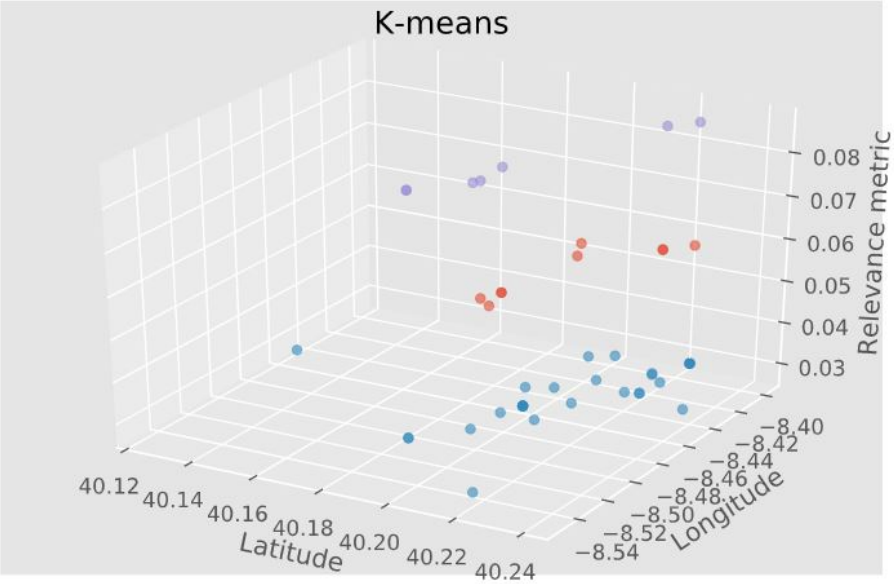


Figure 5.2: 3D scatter plot of the K-means clustering.

5.2.1 Foursquare API

An initial approach utilized the Foursquare API like the state-of-the-art works ([27],[29]). This API is workable in most computer languages and, as expected, has a set of predefined requests that can be made to Foursquare's servers in real time.

The requests follow a simple format of a URL with several completion fields such as $client\_id, client\_secret, version, etc.$ Both the client id and secret require an account to obtain.

To make searches within a specific radius of a coordinate point the necessary inputs are $latitude, longitude, radius$ and $limit$. The first three being self explanatory and $limit$ being the number of points we want the call to return, having a max of 50 POIs at a time. Each of these returned POIs would count as a regular request.

This dataset, however, proved to have several downsides. For one, the restriction on the number requests makes it impossible to conduct a large scale study with real time requests. The way the API returns results further complicates their collection for creation of an offline database. Another main drawback is that useful information for our area classification model, such as ratings and working hours, are considered premium calls and have to be made on each individual POI. Even when making the calls our tests showed that most points lack some of this information. Lastly, Foursquare's data is more oriented towards tourist attractions and leisure POI categories (e.g. restaurants, bars, museums), lacking in less checked-into but equally important categories for inferring mobility habits. All these reasons lead us to search for an alternative solution for collecting the POIs.

5.2.2 Facebook Places API

Having already the experience of working with another social network based API, understanding its negative points, we knew what were the main features that we wanted from a POI data source. Facebook Places dataset proved to be the solution to most of our issues with the previous source. This is mostly due to the fact a dataset had already been constructed for the whole country in a previous work by the AmILab research group [39].

As stated, one of the main attractions of changing to this data source was the fact that we could access the entirety of POIs in the country, in this case Portugal, as it is where our CDR data originated. This meant that we could have an offline database and provided we created efficient methods for searching and accessing data it was the best scenario for our planned algorithms. One more reason to prefer this dataset was the fact that it already contained detailed information on each of the POIs, including number of check-ins and working hours on each day of the week.

In total, the dataset has 221 724 unique points spread over hundreds of categories of different hierarchies. Unfortunately, two main issues required us to create functions to complete the data table.

The first was related to POIs not being classified according to the top level categories, the ones we are mainly interested to use as the possible activity categories in our classification algorithm. For example, restaurants, usually in different categories (e.g. Fast-Food, Portuguese, Asian, etc.) are the same from an user activity point of view. Using a top category level in the Facebook Places hierarchy, $Food\&Beverage$ we group all relevant POIs under the same label, easing classification. The solution to this was to retrieve a separate dictionary file that contained each category name, unique id and the parent category id. With this file at hand and the table of POIs we made a simple recursive function to "walk" the category hierarchy file and fetch the top level categories given a category of any level. When completed, an attribute column was added with this information for each POI.

The second issue was linked to the fact that we intended to use the POIs opening and closing hours to help us in the area classification, and not all points contained this information. Knowing which places were functioning at the time the user was in the area would help narrow the number of possibilities and ideally lead to a better chance at matching the user's activity. To achieve this we created a function (Algorithm 3) that uses points with existing hours information to complete other points of the same category. Here we used the base level categories and not the top ones, as top level are too ambiguous and aggregate POIs that could have very different types of schedules.

Algorithm 3: Complete the missing hours in the POIs dataset.

```
Input: Facebook Places POI Dataset
for each POI in Facebook Places Dataset do
    if hour values exist then
        if workday hours exist then
            Extract all values for opening and closing hours throughout all
             workday days;
            opening = Minimum opening hour;
            closing = Maximum closing hour;
            Replace the existing workday dictionary with the two new values;
        end
        if weekend hours exist then
            Extract all values for opening and closing hours throughout all
             weekend days;
            opening = Minimum opening hour;
            closing = Maximum closing hour;
            Replace the existing weekend dictionary with the two new values;
        end
    end
end
for each category in Facebook Places Dataset do
    if there are any POIs in the category with hour values then
        Make a list of all opening and closing values for the category;
        category opening = mode of all opening hours in category;
        category closing = mode of all closing hours in category;
        if there are any POIs without hour values then
            Add the category opening and closing hours to these POIs ;
        end
    end
end
```

The final result is a table where, other than the regular columns, each point has opening and closing hours for both workdays and weekend as well as top level category (table 5.1). The new column value for opening hours is structured in the following way: [[*workday opening*, *workday closing*], [*weekend opening*, *weekend closing*]].

Table 5.1: Sample of the completed Facebook Places POI table.

name	check-ins	hours	latitude	longitude
Restaurante Aviz	425	[[8, 0], [9, 0]]	39.82468	-7.4915
AZULMIR	15	[[9, 19], [9, 12]]	40.43211	-8.72678
B-Culture	0	[[9, 19], [9, 13]]	41.45011	-8.33808
...	...	...	...	...

category	city	top_category
Portuguese Restaurant	Castelo Branco	Food & Beverage
Wholesale & Supply Store	Mira	Shopping & Retail
Medical & Health	Guimarães	Medical & Health
...	...	...

5.3 Classification of Area Activity

Once we have the users regular locations identified and a complete POIs dataset we reach the final step where all the previous work is combined. To obtain the final output of the classified regular places for each user we need to combine the two types of data in a single function and make some choice regarding the size and shape of the classification area.

Our first possibility was to match routine places at specific hours and dates, with surrounding POIs and run the classification function for thousands or even millions of users individually. The computational time of such a method could possibly take days, or even more, not complying to our goal of creating a scalable methods. Several optimizations were quickly apparent. One, since we are working with the coordinates of the towers there are a limited number of possibilities, with locations repeating between users. Two, as previously mentioned in section 5.1.1, we created a field to divide the CDRs into eight time intervals, and as a consequence we also end up with a limited number of possible time intervals. All this means that we could classify all the locations present in the tower's reference table, at each time interval, for each day type(workday/weekend), and then simply match these classifications to the user's detected routine places, the results of section 5.1.4. Doing the classification this way meant that we still maintain some individuality to the classification, taking into account the day and hours of the visit, while saving a lot in the required computational time.

As computational time is a concern, the first step was to create geometry columns for both the towers reference file and POI dataset, since spatial searches using coordinates are much faster using this data type. For the towers reference we need a geometry polygon that represents the area intended for classification and for the POIs a geometry point of its geographical coordinates. We then ran a search function for each polygon that returned a table of all POIs contained in the polygon.

The next step consisted of filtering the POIs in the area that are open, in each of the eight time intervals defined for routine locations. This seemed trivial at first but finding overlaps in the different time intervals had some challenges as there are several possibilities as well as the change of day when passing midnight. A number of rules ensured that the possibilities illustrated in figure 5.3 are correctly found.

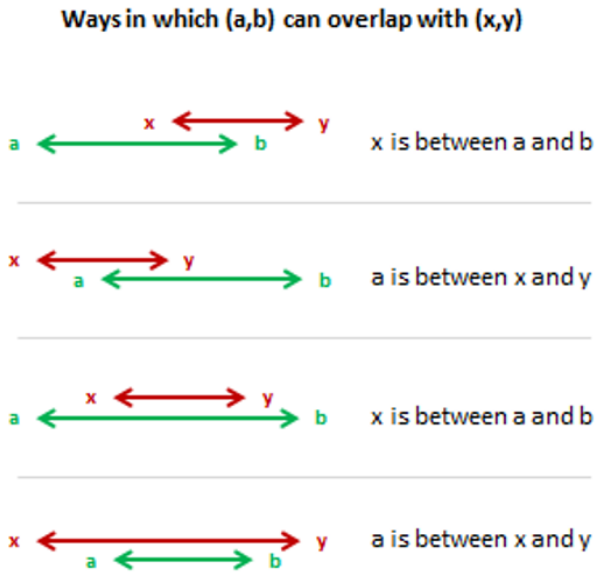


Figure 5.3: Possibilities when matching time intervals.

Let (a, b) symbolize the POI opening and closing hours and (x, y) the CDRs time interval, the base rule is to find the cases in which x is in-between (a, b) and the cases where a is in-between (x, y). We applied the rules presented in figure 5.3 to detect the intersections between the user's records and the POIs opening hours.

After we obtained a table with all points open at a specific time interval we wanted to achieve a classification of most likely activity. As points were already assigned a category by the POI data source we could use those as our class labels. There is a problem however. Lower level categories are sometimes too specific and need to be grouped in broader classes to increase our chance of an accurate classification. For example, if a user visits an area with several types of restaurants (e.g. Fast-Food, Portuguese, Asian, etc.) its very difficult to infer exactly which one he might have visited. If all of then are under the label $Food\&Beverage$, and other categories have a lower prevalence in that area, its safe to say that this classification has a high likelihood of being correct. This is the main reason for using the top-level categories as classification labels.

In a first attempt, we decided to just use the percentage of each category inside the area boundaries. From the area POI table we counted the number of POIs from each label l. Those individual values were then divided by the total number of of POIs in the area a, giving us a new area table with the labels percentage. The label with the highest percentage in the area could be our final classification.

Table 5.2: Example of an area classification table.

top_category_name	pois_percentage	checkins_percentage	c_metric
Food & Beverage	0.125	0.009709	0.001214
Beauty, Cosmetic & Personal Care	0.125	0.019417	0.002427
Religious Organization	0.125	0.064725	0.008091
Medical & Health	0.375	0.284790	0.106796
Shopping & Retail	0.250	0.621359	0.155340
...	...	...	...

However, there was still one more value present in the POIs dataset that we could use in the hope of improving results, the number of check-ins. A check-in is a user registering his presence in the location via a Social Network. As this value is related to the popularity of a given location it could be inserted into the classification as a weight. Based on this improvement the next method of classification used the percentage of label check-ins as an additional value.

So for each label ct in area a we summed the number of check-ins and then divided the individual results by the total number of check-ins in the area a. A new column was then added with this percentage to the existing area table as can be seen in table 5.2. To use both values in the calculation we multiply the label percentage by the label check-ins to obtain our metric for classification. The resulting equation is written bellow in 5.3.

$$C(ct, a) = \frac{Count_{POIS}(ct, a)}{Count_{POIS}(a)} * \frac{Sum_{check-ins}(ct, a)}{Sum_{check-ins}(a)} \tag{5.3}$$

These metrics, although applied in the experimental results, were not subject to evaluation against one another, comparing activity classification. This process of comparative evaluation would require ground-truth data and the prospects of obtaining activity related ground-truth data are outside of this work deadlines. Taking inspiration of works who faced a similar difficulty in obtaining data to validate and evaluate the models ([31],[33],[34]), we compared the results obtained with the known reality of those areas in the results and discussion chapter. This is possible because the study area coincides with the city where the author and other members of AmILab research group have lived for several years.

One issue we haven't addressed yet is the selection of the polygon for the area, this is an important decision as the size and shape of the polygon could greatly influence the final results. Several approaches were considered and are described in detail in the next subsections.

5.3.1 Fixed Radius Approach

By using the Foursquares API, in the first iteration of the algorithms, we ended up using a fixed radius of search around the tower's locations, as this was the default search function to return POIs from the API. This approach had some issues. The fact that a radius is fixed does not take into account that cellular antennas have different signal radius. Although in the project we obtained a value of estimated signal range for each antenna, by that time we had moved on from this approach to the Circular Sections, which will be discussed later in this section.

5.3.2 Voronoi Diagram Approach

Observing the state-of-the-art methods for activity classification we could see that a big majority of these choose to subdivide space into a 500 by 500 meters grid ([30],[31],[33],[35]) as it is considered coarse enough to reduce the noise of classification and detailed enough not to mix different areas.

We believed that Voronoi diagrams could achieve this but with better results, as is seen in the results of the work by Yuan et al. [34]. A Voronoi diagram consists of the partitioning of a plane with n points into convex polygons such that each polygon contains exactly one generating point and every point in a given polygon is closer to its generating point than to any other. A Voronoi diagram is sometimes also known as a Dirichlet tessellation. The cells are called Dirichlet regions, Thiessen polytopes, or Voronoi polygons. To create workable polygons from a Voronoi diagram we resorted to a combination of two python libraries, scipy and shapely giving the tower locations as input. An example of the diagram created from the tower's reference file can be seen in figure 5.4.

An approach such as this seemed like a good improvement over the fixed radius. Taking the area closer to each of the points in the reference file and finding all contained POIs for obtaining a classification. Even so, this approach was actually based on some wrong premises. As explained in section 2.1 it is wrong to assume that a user always connects to the closest cellular antenna as this depends on several factors. So by having an event in their CDRs with a certain *cellid* does not necessarily mean that they are closer to that identified antenna than to any other and the area classification with this method could lead to results distant from reality.

5.3.3 Circular Sections Approach

Finally we reach the last representation for the area the user could be when connected to a certain antenna to make an event.

Through our fortnight meetings and presentations with the telecommunication service provider it was discussed the possibility of obtaining new information as part of the tower's reference file. These advances in information would give us the angle and range values of all antennas. With such values we could create a polygon of our own, one for each antenna, that accurately represented the area a user needs to be in order to be assigned to that antenna.

Figure 5.4: Voronoi Diagram created from the tower's locations.

To create the polygons we used an existing python library, shapely, which has the exact tools needed already implemented in readily functions. The final area polygons are created by the intersection of two other polygons. One is a circular buffer with center at the antenna and a radius equal to the estimated signal range. The other is a circular sector with initial and final angles obtained from the antennas attributes. This is done because the existing function in the library for creating circular sectors is not capable of receiving a value in meters, such as the radius. The easiest way was then to create a circular buffer with a different projection and convert it to the WGS84 datum, intersecting the circular sector and creating the final polygon. An example of an antenna circular sector can be seen in figure 5.5.

Now an area exists that, to the best of our knowledge, indicates the area the user was when a CDR event is made. Although antenna's areas vary to a great extent and some have a very large dimension, the expected increment in certainty regarding areas visited by the user could be worth these downsides. Furthermore this method could not be found in works with similar goals at that fact alone made it worth exploring further.

5.4 Improving the User's Location

This part of the work originated from the search of the best possible area for classification. As explained in the section 5.3.3, changes in the towers reference file allowed us to create polygons directly representing the area of each antenna. In this process an idea appeared to improve the position associated with the user when connected to an antenna.

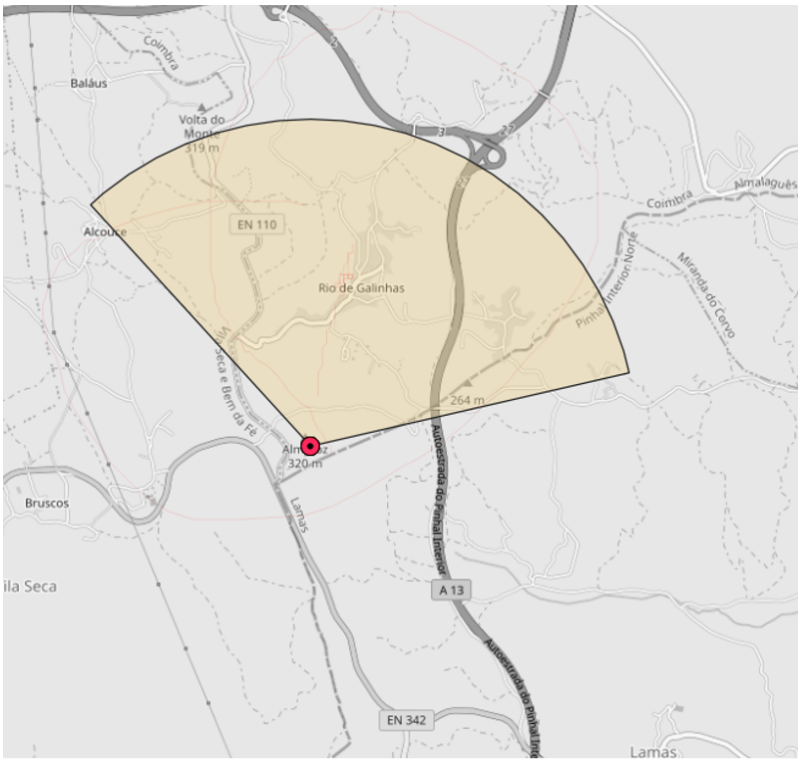


Figure 5.5: Example of a antenna's area.

Throughout this work until now we have been using the locations of cellular towers. Even though CDRs identify the ID of an individual antenna within the tower, all antennas belonging to the same tower were bound to its location. This implies heavy limitations in the accuracy of trying to position a user from its CDRs. For example, (see figure 5.6) in a tower whose cells have a 5 kilometer radius from the origin point, an event made in positions a and b would always be categorized as the origin point o, even though theoretically the user could be in positions that differ a total of 10 kilometers between them.

With the new values for the angle and range of antennas we are able to know, not only the rough direction from the origin point where the event was made but also the maximum distance at which it could be made. If capitalized, this extra knowledge could increase the accuracy of previous methods that use user's coordinates from CDR tables, such as detection of home, workplace and other routine places as well as improve future research projects.

As we were not interested in using the whole area polygon as the new position we needed a new spatial point. Once the polygon of the antenna's area is created, a new coordinate point can be used to replace the tower's coordinates when an event is made from that antenna. Previous methods could be easily adapted if the tower's reference table, present in table 4.3, was updated with new values for each individual antenna.

Figure 5.6: Illustration of antennas signal area.

Our first idea used the centroid of the antenna's area polygon, the arithmetic mean position of all points in the polygon (figure 5.7(b)). This averaged the possible positions inside the polygon and gave the best compromise. An alternative idea had its inception in a clarification's meeting with the client. It was discussed that the further away the client is from the tower, the lower is the antenna's signal, resulting in an increased probability of connecting to another antenna. With this in mind, we also applied a positioning method that placed the new coordinates point closer to the tower, choosing to use the middle point between the tower and the area centroid (figure 5.7(c)). An example of all versions for the same CDR event, including the original positioning, can be seen in figure 5.7. Both position versions were compared to the original tower's coordinates, in terms of accuracy in detecting the user's home location. This comparison analysis is present in chapter 6.

(a) Location version 1

(b) Location version 2

(c) Location version 3

Figure 5.7: Visualization of the three positioning methods.

Chapter 6

Results and Discussion

This chapter contains a detailed description of the results directly obtained from applying the outlined methodology in order to answer the objectives posed in chapter 1. In the first section, experimental results have been summarised and discussed for each research point individually. In the second part, some results have been validated with resource to ground-truth data. Qualitative evaluations are also carried out in cases where ground-truth was not obtained. Appropriate tables and graphs have been used to describe the findings and improve the discussion of used methods. Our analysis, as previously mentioned, is focused in the district of Coimbra, Portugal.

6.1 Experimental Results

In this section, we report experiments conducted to test the behavior of the proposed methodology. The main results have been discussed individually, presenting the output and discussing how they were used to answer the research question and final objective of this work - The identification and classification of places of interest using anonymised mobile communication data.

6.1.1 Load Sharing

Preparing the data tables is an important factor to obtain the best results possible out of the implemented methods that detect location patterns. The exclusion of anomalous records, such as the ones identified as network Load Sharing was a frequent point present in state-of-the-art works ([20], [21], [22]).

Running the method defined in section 5.1.2 to flag and remove the presence of the Load sharing effect in the dataset we encountered a total of 5798 anomalous pairs of records. This amounts to approximately 0.0274% of the events present in the dataset.

Table 6.1: Load Sharing Results.

Load Sharing count	5798
Records percentage	0.0274%
Mean speed	240.26
Std	512.80
Min	55.05
25%	75.09
50%	114.67
75%	200.41
Max	7892.75

As per the work of [22], the speed threshold at which an antenna reselection was flagged as Load Sharing was set at 200 km/h (kilometers per hour) or 55.5 m/s (meters per second). Analysing the records pairs that were above the threshold we obtain the statistics in table 6.1 where the speed values present are in meters per second. With an average speed value of 407 m/s these are far from possible movements in a normal transportation mode and as such we deemed them as correctly identified. A pair of records with identified Load Sharing can be seen as example in figure 6.1. These two records had a time difference of 5 seconds between them with a calculated speed of about 840 m/s.

```
----------------------------------
0:00:05 840.0542890806266
Unnamed: 0                                                256157
dtime                                          2020-09-10 16:09:57
cellid                                                    490262
MSISDN       033ECE58CEC49373EF46326D840458D7744388109547FE...
latitude                                                41.203019
longitude                                               -8.562231
Name: 252344, dtype: object
Unnamed: 0                                                256162
dtime                                          2020-09-10 16:10:02
cellid                                                     30078
MSISDN       033ECE58CEC49373EF46326D840458D7744388109547FE...
latitude                                                41.180473
longitude                                               -8.531594
Name: 252349, dtype: object
----------------------------------
```

Figure 6.1: Example of a Load Sharing detection in the dataset.

6.1.2 Home and Workplace Locations

Undoubtedly the most important places for the majority of the population are their home and workplace. These are, in fact, places that amount to a large portion of time spent throughout our lives. Home, in particular, gained increased relevance in recent times. Due to a global pandemic, people were forced to spend more time at home and even adapting the space for other activities, like work, to be made all from the same location. Since they are prominent places in people's lives, they fall inside our objective of detecting meaningful and regular ones.

Detecting the most probable home and workplace of each user was the first pattern inference method in the defined methodology. For this we clustered CDR locations inside a specific time interval attributed with being in each of those two places. The interval from 7PM to 9AM was assigned to home and the 9AM to 5PM interval to workplace. We are aware that the chosen schedules are fixed and do not accurately capture irregular behaviours, like students or night workers. We felt like these cases were not in our best interest to pursue, in the interest of not loosing focus of our main goals and because of the project time constraints. Furthermore, there was already a thesis hosted by AmILab, parallel to this one, whose primary objective was to enhance home and workplace location detection, including making the distinction between those mentioned cases. Results from both works could eventually mutually improve each other in a follow up research.

This step was done primarily in our work to improve the identification of other routine places without wanting to wrongly detect and mislabel home and workplace as other activities. However, this section actually received increased importance in the validation phase, since the only ground-truth data made available by the telecommunications service provider was for home location. A small dataset, comprised of a sample population in the original CDRs file, was annotated for home postal code coordinates and used in the upcoming section 6.2.1 to validate our approach.

Its important to refer the output of the implemented home and workplace detection method. Two tables were created containing the *msisdn*, unique identifier of each of the users, as well a pair of latitude and longitude coordinates. These coordinates where obtained by calculating the centroid of the most active antenna's cluster in the respective hours for home or workplace. Figure 6.2 shows examples of both places visualized for randomly selected users. Users are assigned a unique color and their homes and workplaces identified by the "h" and "w" letters respectively.

Antennas belonging to the most active cluster in both time intervals, as previously explained, are excluded from the dataset prior to running the following methodology for identifying other routine locations.

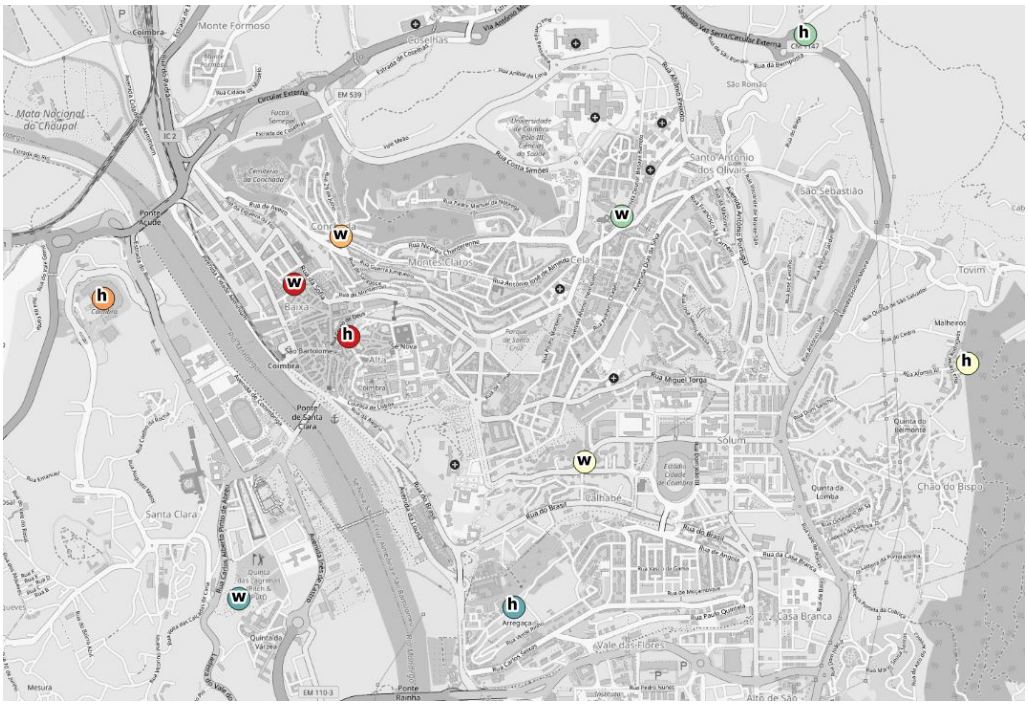


Figure 6.2: Detected Homes and Workplaces for random user in the city of Coimbra.

6.1.3 Other Routine Locations

According to the paper by C.Quadri et al. [27] people can be divided into two basic types: returners, those who are very regular in their daily mobility; and explorers, those who are inclined to break out of their daily mobility routine and explore new places [27]. While their work focused on people's exploration related behaviour, out of their regular paths, ours focused on the behaviour of moving back and forth between a relatively small set of places.

We started with the most regular set of places possible, home and work, and then, moved on to the remaining points that fundamentally capture the mobility of the individuals. To do so, we once again made use of the locations present in CDRs and found patterns of recurrence in these locations. By the techniques outlined in section 5.1.4, we calculated a relevance metric for each individual location and input the results to the partition-based K-means clustering algorithm. The K-means divides the locations into three distinct groups, using solely the relevance metric. The group containing the highest metric values are considered the Most Visited Places (MVPs) with the second highest being attributed to places that are also visited with frequency but not as much as the first, Occasionally Visited Places (OVPs) and then finally are the Exceptionally Visited Places (EVPs) associated with exploration behaviour. We collected the two groups with highest relevance metric, MVPs and OVPs, discarding the remaining ones (EVPs). It is worth to recall that each detected routine place has associated time-interval and workday/weekend values. There is a possibility that a location is identified more than once, if the user is frequently found there in different hours or different days, as areas can have different purposes depending on these factors.

For every user, just like it was for the home location, a table entry was created with the identified MVPs and OVPs coordinates. The table also included a column with the calculated relevance of each place to the respective user.

These places can be visualized in a map layer like the example in figure 6.3, where a random user was selected with and his MVPs and OVPs plotted. The figure shows the main urban center of Coimbra with the user's detected MVPs in green and OVPs in yellow.

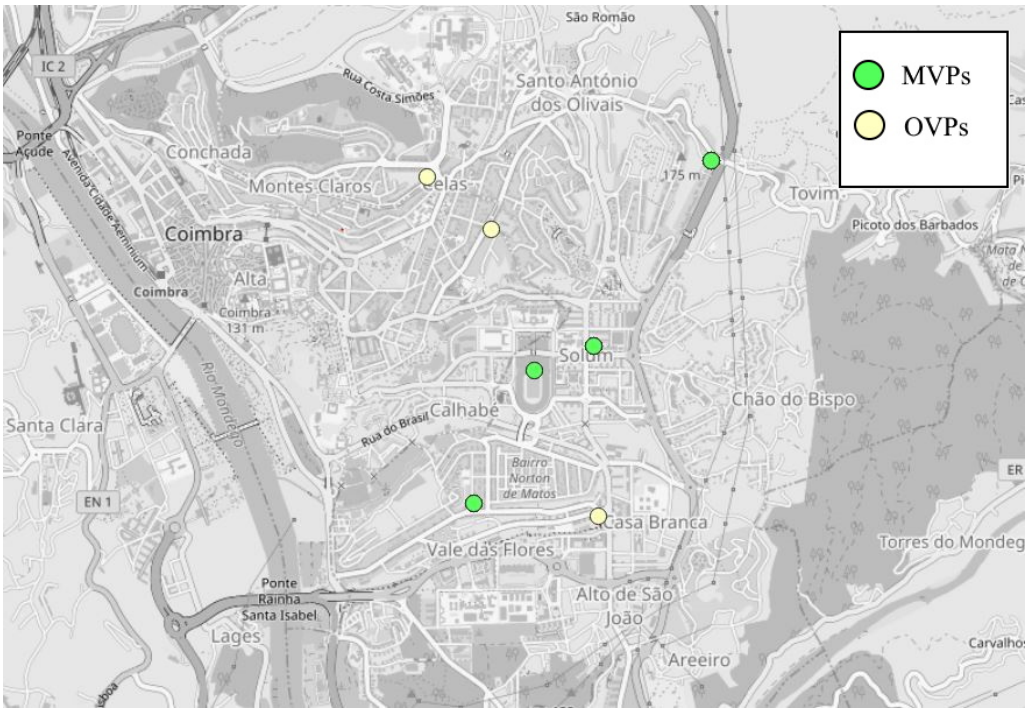


Figure 6.3: Detected MVPs and OVPs for one user.

This set of places are inherently more difficult to detect, as there likely is a comparatively reduced number of events made in each of them, and for some routine places, the user might not make any event at all. We expect the result to not be as accurate as of the home location, but to be nonetheless enhanced by using the approach for location improvement (section 6.2.2). The goal behind finding these locations was to ultimately classify each of them according to the most likely activity. The area classification results, to be presented in the next section, give continuity to this point.

As of the moment of writing this thesis, no validation data could be attained to test our approach in identifying these places. The privacy concerns associated with making a questionnaire to a sample of our study population proved to be too big an obstacle for the available time frame. However, there is a possibility for this data to become available in the continuation of this project that involves both AmILab and the telecommunications service provider.

Table 6.2: Fixed Radius Area Classification table.

Workdays		Weekends	
Temporal interval	Classification	Temporal interval	Classification
(0, 3)	Food & Beverage	(0, 3)	Food & Beverage
(3, 6)	Arts & Entertainment	(3, 6)	Arts & Entertainment
(6, 9)	Sports & Recreation	(6, 9)	Sports & Recreation
(9, 12)	Sports & Recreation	(9, 12)	Sports & Recreation
(12, 14)	Sports & Recreation	(12, 14)	Sports & Recreation
(14, 17)	Sports & Recreation	(14, 17)	Sports & Recreation
(17, 20)	Sports & Recreation	(17, 20)	Sports & Recreation
(20, 24)	Sports & Recreation	(20, 24)	Sports & Recreation

6.1.4 Area classification

Getting closer to the ultimate goal of classifying the user's routine places we needed a method that could, given a location, probe the surrounding area for points that could be related to the user's mobility. These points and their respective attributes, attained through a specific POI dataset, would then be accounted for in inferring the most likely activity for the user at the location.

The followed approach had the opening hours of each POI taken into consideration for classification, as well as the distinction between workdays and weekends. For each combination of time interval (defined in section 5.1.1) and day type, only the POIs open in that schedule are used for classification.

The first method implemented was the fixed radius search. Centered around every tower, a circular buffer of 500m was created. Every POI whose position fell within the buffer area was identified and accounted for in inferring the most likely activity. Figure 6.4 represents an example of the circular fixed radius buffer for a selected location. The center antenna is identified by *cellid* number 21063 and this same antenna will appear in subsequent examples of area classification to illustrate the differences that the area's shape and size have on the results. This antenna is located at a tower in the left margin of the Mondego river in the city of Coimbra. It is surrounded by several interesting points including some famous tourist attractions as well as street commerce and restaurants. In Figure 6.4, POIs are categorized by their base level categories, however in this approach, as described in section 5.3, only top level categories are used as classification labels.

Table 6.2 presents the results of classification for the fixed radius approach. The example corresponds to the area and POIs visualized in figure 6.4. Temporal intervals in this table have the following format, $(starting\ hour, finish\ hour)$, using a twenty four hour notation.

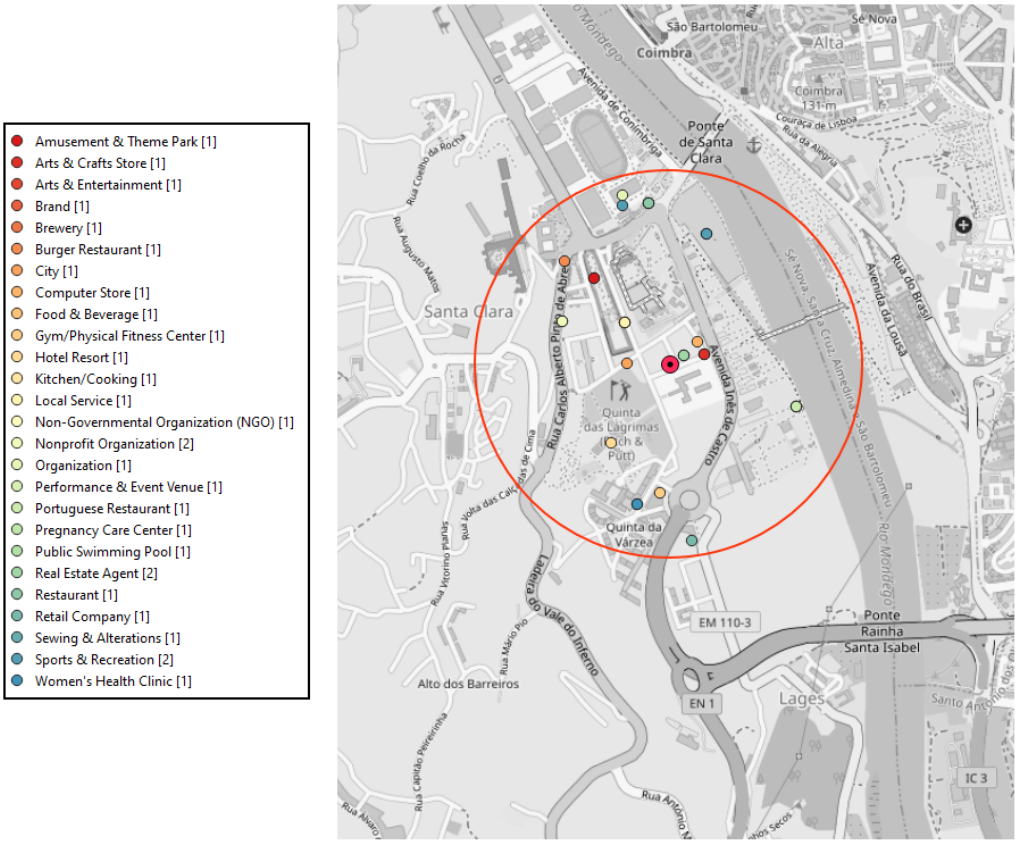


Figure 6.4: Fixed Radius Area Classification example.

Succeeding the fixed radius approach, addressing its flaws, we needed a way to dynamically create areas with different sizes. With the addition of the individual antenna's radius to the reference dataset, it was confirmed that antennas signal ranges have a large disparity of values, as analysed in section 4.2.2. This disparity causes the fixed radius approach to not be ideal in terms of area selection. Our proposed solution was to switch the classifications areas to ones obtained by a Voronoi Diagram, created with the antennas coordinate points. Voronoi cells have arbitrary shape and size depending on the position and density of the generator points given as input. This is an improvement over the first approach as more isolated antennas have a larger area for classification and antennas in high density areas have a proportional smaller area.

The same antenna, *cellid* number 21063, appears once again in figure 6.5 as the generator point for a Voronoi cell. As before, the POIs whose location is within the boundaries are displayed and color coded according to their base level categories. This area, thanks to its larger shape, contains the same points found in the first approach with the addition of some new ones, possibly leading to new activity classifications.

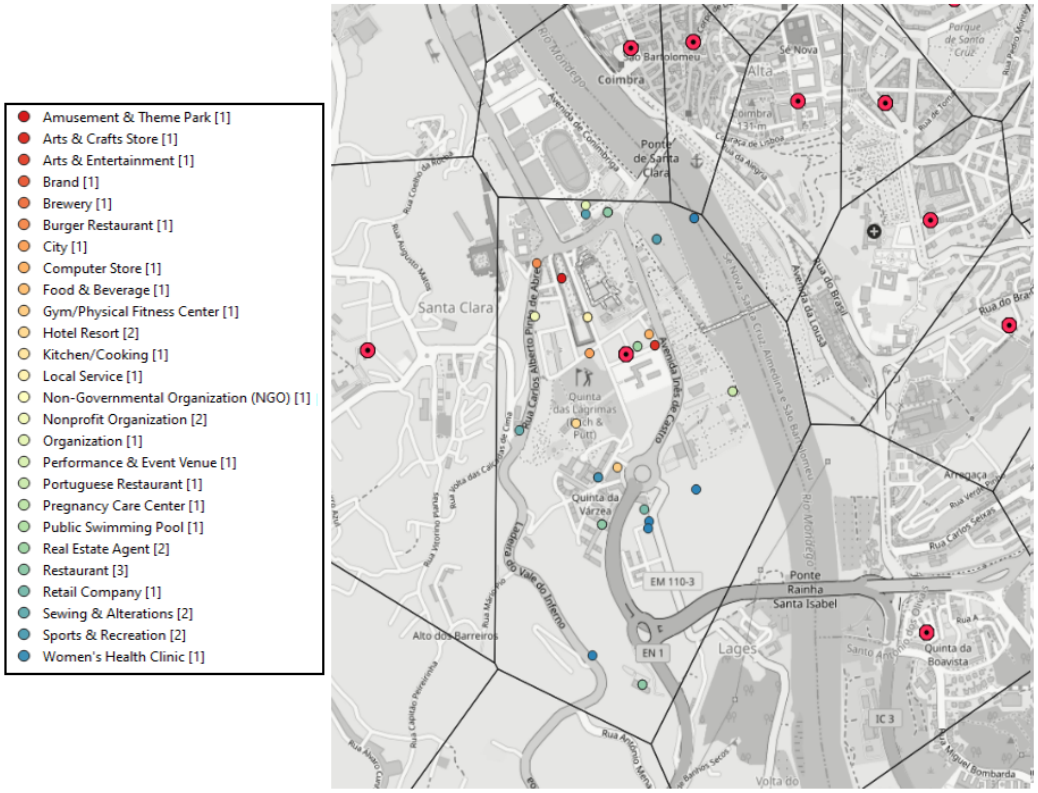


Figure 6.5: Voronoi Cell Area Classification example.

In Table 6.3, we can observe that the change in area had an impact on the classification of some of the time intervals. Although the most probable activity remains unchanged in the intervals between midnight and 6AM, the remaining ones obtained a different classification. The extension in the area's shape meant that more restaurants are now inside the boundaries. The increased prevalence of this category seems to have influenced the results in favour of *Food & Beverage* for all of the time intervals, with the exception of 3AM to 6AM when no restaurants were found open.

Even though the Voronoi Diagram solved our problem of dynamic areas, we later understood that it had some fundamental flaws. Voronoi cells represent all the points that are closer to the contained generating point than to any other generating points. This approach assumed that a user connected to an antenna is closer to it that to any other, which is true in some cases, but not all. For example, certain obstacles like landscape or buildings can force the connection to an antenna that, even though is further away, might be at a better angle for transmitting the signal.

Table 6.3: Voronoi Cell Area Classification table.

Workdays		Weekends	
Temporal interval	Classification	Temporal interval	Classification
(0, 3)	Food & Beverage	(0, 3)	Food & Beverage
(3, 6)	Arts & Entertainment	(3, 6)	Arts & Entertainment
(6, 9)	Sports & Recreation	(6, 9)	Sports & Recreation
(9, 12)	Food & Beverage	(9, 12)	Food & Beverage
(12, 14)	Food & Beverage	(12, 14)	Food & Beverage
(14, 17)	Food & Beverage	(14, 17)	Food & Beverage
(17, 20)	Food & Beverage	(17, 20)	Food & Beverage
(20, 24)	Food & Beverage	(20, 24)	Food & Beverage

The final approach originated when we knew it was possible for us to obtain precise data that allowed to estimate the signal area of antenna. This not only would solve the inherent problems of the fixed radius method but the Voronoi cells as well. With the direction and range values we could have a good representation of the area the user was, in order to connect to the corresponding antenna. This allowed the boundaries and contained POIs to be more precise and relevant to the real user location. The newly created areas had the shape of a circular sector with the center point of the circle being the antenna's coordinates. As far as we know, this approach has not been replicated by any previous works in literature and is a new attempt for enhancing state-of-the-art activity classification.

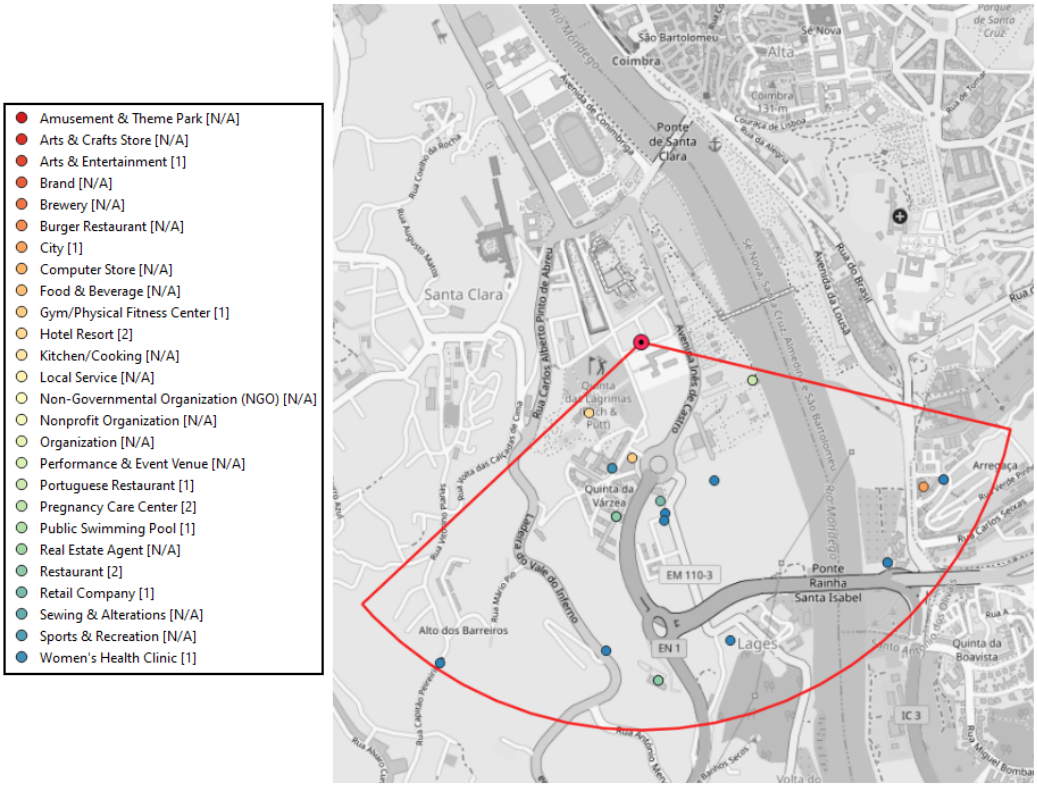


Figure 6.6: Antenna Cell Area Classification example.

Table 6.4: Antenna Signal Area Classification table.

Workdays		Weekends	
Temporal interval	Classification	Temporal interval	Classification
(0, 3)	Food & Beverage	(0, 3)	Food & Beverage
(3, 6)	None	(3, 6)	None
(6, 9)	Sports & Recreation	(6, 9)	Sports & Recreation
(9, 12)	Medical & Health	(9, 12)	Medical & Health
(12, 14)	Medical & Health	(12, 14)	Medical & Health
(14, 17)	Medical & Health	(14, 17)	Hotel & Lodging
(17, 20)	Medical & Health	(17, 20)	Hotel & Lodging
(20, 24)	Sports & Recreation	(20, 24)	Sports & Recreation

Figure 6.6 continues the usage of the same *cellid* as previously seen in the other area examples. This time the defined area for this antenna is kept only in a certain direction, extending and capturing zones previously not attributed to the antenna. Results in table 6.4 confirm that idea. For most of the time intervals, results were vastly divergent from the preceding area approaches. Points of Interest belonging to the $Medical\ \&\ Health$ category are now the most notable between 9AM and 8PM. Additionally, we can finally see an example of disparity between workday and weekend classifications in the time intervals between 2PM and 8PM.

Throughout these multiple approaches we believe to be improving classifications, bringing results closer to the real activities users might have taken part in those places. However, without ground-truth data there is no real way of knowing if any approach is better than the remaining. Due to the primacy of privacy, annotated activity data would have to be collected and delivered to us by the telecommunication service provider. Only they possess the real identities of the users in the CDRs file, so only they could match possible questionnaire's or annotations with the *msisdn* identifiers in the records table. Despite our best efforts to push this request, data for validation was not obtained in time for the deadline of this thesis. We will, however, use a discussion approach found in state-of-the-art works with similar difficulties ([31],[33],[34]), using knowledge from the study area to make a qualitative analysis of the results.

In terms of qualitative verification, we consider some typical functional areas of the city of Coimbra and discuss their corresponding classification according to our local knowledge of the study area. Like previously stated, annotated ground-truth data was not a possibility within the deadlines of this thesis.

We took as first example the antenna visualized in figure 6.6 and classified according to table 6.4 for discussion. The predominant activity classification throughout workdays is *Medical & Health* and *Sports & Recreation*. This is due to a high number of POIs of these top level category open in the area, including care centers, health clinics, a physical fitness center and some outdoor Padel fields, amounting for a higher prevalence than any other top level categories. This area also intersects the location of some popular restaurants, but is only classified as *Food & Beverage* in the time interval past midnight until 3AM, when all other categories have already closed down. This is consistent with reality since POIs related to *Food & Beverage*, which include bars, cafes and restaurants, tend to close at a later hour than health or sports related businesses. Also consistent with known reality is the fact that Gyms and Sport activities open sooner that other types of POIs, for people that prefer to take part in these activities early in the morning, giving strength to the 6AM to 9AM classification. The last point to focus in this area is the change in activity at the weekends between 2PM and 8PM. The explanation could be related to the presence of the famous Hotel Quinta das Lágrimas, whose gardens are a popular visit location and might be a relevant choice for a walk during the weekends.

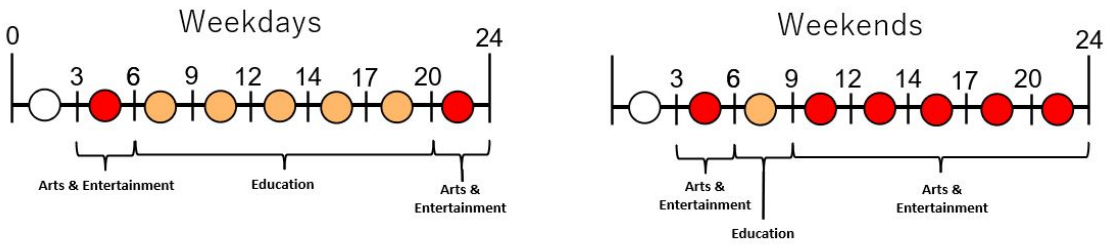


Figure 6.7: Classification for cellid number 609067.

Another example chosen for this analysis is *cellid* number 609067, a cell with a relatively small signal area that intersects the Pólo I, the original part of the University of Coimbra, as well as the city's botanical garden. Classification results for this area can be seen in figure 6.7 with the area itself illustrated in figure 6.8. Our expectations for this cell, by our personal knowledge of the zone, seemed to be correct. During workdays the vast majority of time intervals are found to be the *Education* activity, matching with normal university lecture schedules. In the weekends, although still present, as seen by the classification between 6AM and 9AM, education related POIs are mostly closed, giving the prevalence to *Arts & Entertainment*.

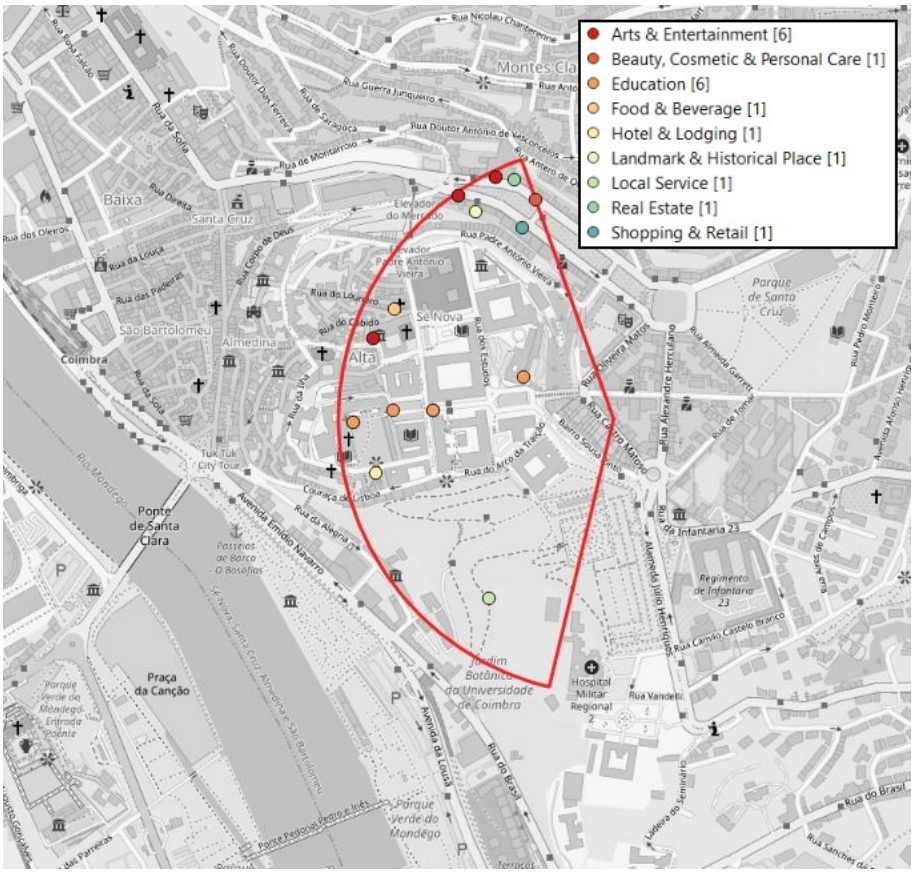
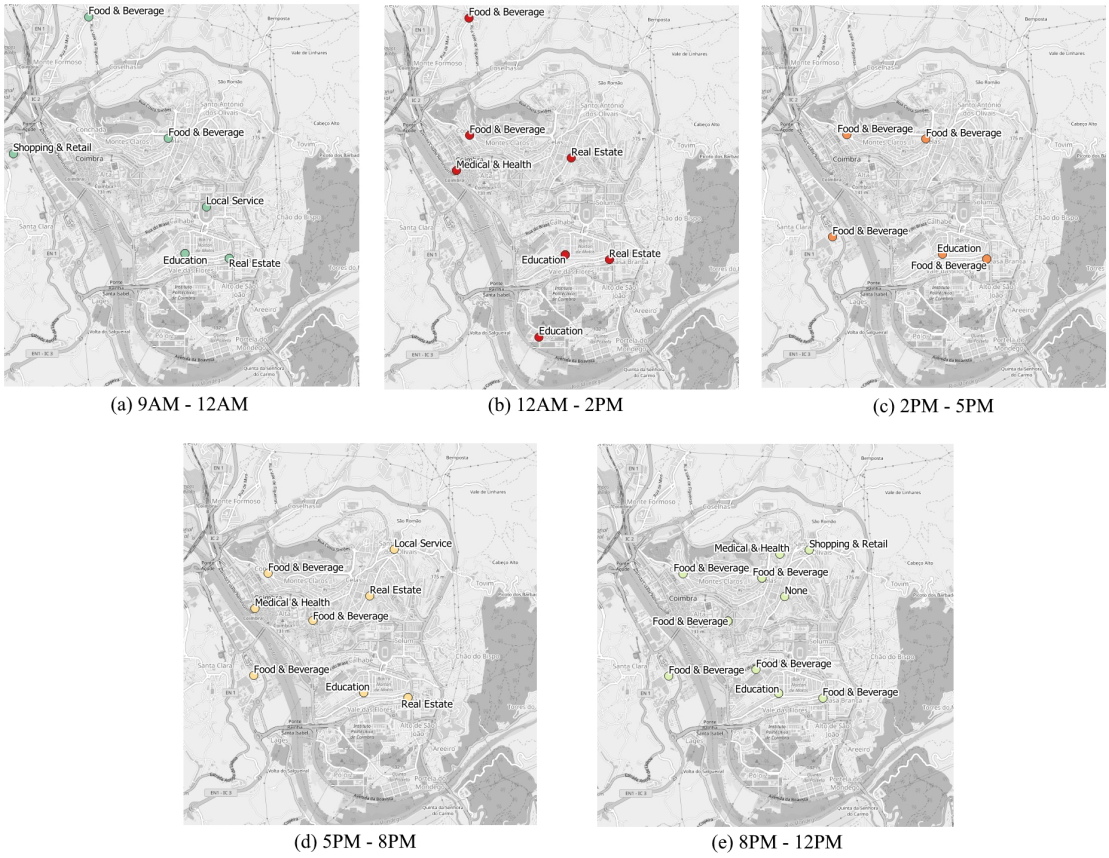


Figure 6.8: Antenna area for cellid number 609067.

6.1.5 Routine Places Classification

It is important to reiterate that area classification is done to all antennas present in the dataset. Only after having the areas classified, including all time intervals and day type possibilities, are they matched with each one of the user's routine places. The matching key is the corresponding antenna identifier, *cellid*, that is present in both the output for area classification and routine place detection.

This allows us, for each user to obtain a visualization similar to the one present in figure 6.9, where routine locations are separated by time interval and labeled according to the prevalent activity. Some locations might repeat for different time intervals but have a different classification, as we took into account the opening hours of POIs. These routines are inferred from 2 months of activity, between September and October of 2020.



(a) 9AM - 12AM (b) 12AM - 2PM (c) 2PM - 5PM

(d) 5PM - 8PM (e) 8PM - 12PM

Figure 6.9: Routine locations activity by time interval.

Table 6.5: Routine location activities table.

Workdays		Weekends	
Temporal interval	Activities	Temporal interval	Activities
(9, 12)	[Local Service, Real Estate, Education, Food & Beverage]	(9, 12)	[Real Estate, Shopping & Retail, Food & Beverage]
(12, 14)	[Education, Real Estate, Medical & Health, Food & Beverage]	(12, 14)	[Food & Beverage]
(14, 17)	[Beauty, Cosmetic & Personal Care, Real Estate, Education, Food & Beverage]	(14, 17)	[Food & Beverage]
(17, 20)	[Real Estate, Education, Food & Beverage, Medical & Health, Local Service]	(17, 20)	[Food & Beverage]
(20, 24)	[Food & Beverage, Education, Medical & Health, Shopping & Retail]	(20, 24)	[Food & Beverage]

Another way to visualize the activity patterns of the user in by grouping by the type of day and time interval, as demonstrated in table 6.5. This table shows in each time interval, for both workdays and weekends what were the identified activities according to his routine places. The user used for this example is the same one as figure 6.9. This data would obviously benefit from validation using either a user questionnaire or volunteer's GPS data with annotated check-ins. Although we worked towards obtaining such data, in the end it was not a possibility during the timeframe of the project.

6.2 Validation and Evaluation

As research progressed, the attainment of ground-truth data was increasingly a concern, as it is a crucial part for validating our algorithms as well as to allow for improvements to be made by comparing the results of each method. Ground-truth data attainment is especially a challenge in a field whose research is centered on information considered private and sensitive as is the case of call detail records. Only by having direct contact with a telecommunications service provider throughout this thesis were we able to secure a validation dataset. This allowed us to to verify that some approaches were performing as expected, in line with their objectives.

6.2.1 Home Location

As previously mentioned, the only validation dataset that we could obtain was the home location for a sample of the study's population. For this kind of sensitive data collection the users identification is anonymised before being provided for research purposes, just like in CDRs, making it impossible for anyone outside of the data provider to connect the information in the ground-truth with real people.

The data table obtained for the first home validation was composed of 4628 users (13% of the CDRs users) home locations, with the corresponding information for each comprising a single row. Columns fields contained two identifiers (*msisdn* and *cod_unico*), the shared keys to the CDRs table, as well as the type of tariff , $dsc\_familia\_tarifario$, and postal code with the corresponding centroid coordinates ($cod\_postal\_utiliz1$, *loncentroide* and *latcentroide*). The postal code centroids are calculated, or obtained, by the telecommunication service provider and made available to us in Cartesian coordinates. A sample of this file can be seen in table 6.6.

To validate the results obtained by applying the methodology presented in section 5.1.3 we used the subdivision of antennas by urban and rural areas as done in the Exploratory Data Analysis area of the results (section 4.2.2) inspired by the work of [37]. As shown by our analysis of towers, there is a large disparity between the average radius of antennas in more dense urban regions and less dense rural regions. After having attributed to each antenna a field containing one of these classes, we calculated two T threshold values, one for each class by averaging the antenna's signal range.

Also following the approach of Mamei M. et al. [37], we calculated the distance from the detected home centroid to the ground-truth coordinates using the geodesic distance between two points, just like we had done in section 5.1.2 to detect load sharing. Finally we compared the obtained distance values with the average radius for urban or rural areas. If the distance value between the detected place and the ground-truth was below our threshold T , we considered to have found home location correctly and score 1 positive identification. All the identified homes that are above the threshold are considered as being the wrong location. For both classes we divided the positive identifications by the total number of user's in the class and obtain our accuracy values.

Table 6.6: Sample of the home validation table

msisdn	cod_unico	dsc_familia _tarifario	cod_postal_utiliz1	lon centroide	lat centroide
0003F54...	5C60290...	M Movel Pos-Pagos	9125-239	-16.8316	32.6543
000E46D...	0F708E7...	Unlimited	3030-871	-8.3918	40.1723
00344D9...	5A61CB7...	Unlimited	3030-2185	-8.4057	40.1955
003F030...	9D555E7...	M Movel Pos-Pagos	3020-227	-8.4119	40.2347
...	...	...	...	...	...

Table 6.7: Home validation results

	Acc. Urban	Acc. Rural	Mean Distance	Std.	Q1	Q2	Q3
No filter	65.10%	29.32%	40.94	149.7	0.710	1.898	9.160
<50 km	66.55%	31.25%	4.509	8.747	0.600	1.384	3.432
<35 km	68.34%	42.04%	3.254	5.552	0.580	1.306	3.115
<20 km	68.99%	47.60%	2.424	3.224	0.570	1.255	2.807

Analyzing our initial results, we encountered some irregular values regarding the distance between detected and real homes. Some users had calculated values in excess of a thousand kilometers between both points. Even though homes were identified in the Coimbra district, since user's were selected for data collection by having a large percentage of events in this area, real homes were spread out all over the country, including areas like the Algarve (Figure 6.10(a)) and even the Madeira and Azores archipelagos (Figure 6.10(b)). These findings were discussed with the client and data provider. The feedback we got implied that the majority of these values would be related to the student population in this city, a large portion of which live away from their main residence area. These were agreed to be classified as annotation errors, as those users main addresses are not the ones they spend most of their time in. To remove these annotation errors several threshold distance were defined and tested for values of (20, 35 and 50 kilometers), effectively separating outliers from the rest of the results.

Figure 6.10: Examples of irregular detected d and real r home locations

Using the validation algorithm here outlined we obtained the initial results contained in the rows of table 6.7, each row having a different outlier threshold. Besides accuracy, the tables also contain information on the mean, standard deviation and quantiles of the distance between detected and real locations. All values, except for accuracy, are displayed in kilometers. As it could be observed by the results, filtering outliers has a great impact in all performance metrics. One thing to take into consideration is that distances are measured in a straight line, and navigating between these points in a road network could increase that value. We choose to continue using the 50km threshold for the following analysis as we believe that distances above this value have a high chances of being an annotation error.

The following step was to backtrack to the DBSCAN input parameters and try to improve the results by optimizing these values. The process was possible by this point as we now had the accuracy and other metrics as guidance. As described in the DBSCAN Algorithm Section, 2.2.3, there are two core inputs, the neighborhood size in terms of distance (Eps), and the minimum number of points in a neighborhood for its inclusion in a cluster (MinPts). As we are working with points in Cartesian Degrees, the Eps value is also in the same unit. Optimizing these values is a balancing act, as for example, a low value for Eps might not cluster all the antennas that intersect the user's home and a value two high might add unwanted antennas to this cluster, distancing the detected home from the real one. Table 6.8 shows the various runs made with different combinations of values.

As inferred by the tests in table 6.8, varying the Eps had little to no effect, while increasing MinPts seemed to have a positive impact in all values. Thereby, increasing the number of points necessary in the neighborhood for a point to be included in a cluster is more relevant for detecting home location.

Table 6.8: Tests results varying the *Eps* and *MinPts* parameters

	Acc. Urban	Acc. Rural	Mean Distance	Std.	Q1	Q2	Q3
Eps=0.005 MinPts=2	66.55%	31.25%	4.509	8.747	0.600	1.384	3.432
Eps=0.005 MinPts=5	69.12%	49.41%	4.379	8.640	0.574	1.348	3.285
Eps=0.005 MinPts=8	70.73%	55.73%	3.857	8.095	0.473	1.160	2.941
Eps=0.007 MinPts=2	69.06%	47.66%	4.509	8.749	0.601	1.380	3.421
Eps=0.007 MinPts=5	69.25%	49.76%	4.374	8.639	0.574	1.336	3.260
Eps=0.007 MinPts=8	70.99%	55.48%	3.849	8.070	0.481	1.160	2.939
Eps=0.009 MinPts=2	68.97%	47.89%	4.502	8.748	0.604	1.389	3.381
Eps=0.009 MinPts=5	69.22%	49.53%	4.377	8.641	0.589	1.354	3.260
Eps=0.009 MinPts=8	70.57%	55.43%	3.880	8.096	0.4907	1.193	2.967

We obtained a best value of 70.99% accuracy for urban and 55.73% for rural areas using the implemented methods. Comparatively to other works, like [20], that obtained values of 86% (with a distance threshold of more than 4000m), our results seem slightly worse, however we keep the distance threshold small compared to their results. Our thresholds of success are in fact the average radius of antennas, 2084m and 2961m for urban and non-urban/rural areas respectively. If theoretically we had increased our distance threshold for success, accuracy values would potentially have increased accordingly. The work of [37] also presented higher values (90% precision), however their annotated dataset only contained 11 locations.

6.2.2 Location Improvements

In this work, we initially intended to detect and classify routine places using CDRs. One of the known challenges of working with this data type is the associated uncertainty in the position, something that several state-of-the art authors focused on improving as a end goal or as a way to enhance their study results ([14], [25]). Even though it was not in our predefined path, by the advances made throughout the search for an optimum classification area, an opportunity emerged to try and improve the locations present in the CDRs, and in doing so benefiting all other implemented methods.

As presented in Section 5.4, we calculated a new user position dependent on the antenna that is receiving the connection signal from the mobile device. To our understanding, and by research done in literature, this is a new approach at improving CDRs underlying spatial uncertainty. Although we did not have a specific position validation dataset to test and compare our methods coordinates with the original records, we used the home validation dataset. The implementation of improved positions should affect the results of the home detection algorithm, since all positions in the records are changed to new ones. We used the validation algorithm for the home location described in section 6.2.1, running different position versions, to observe and discuss if there are increments in accuracy related to improvements in positioning.

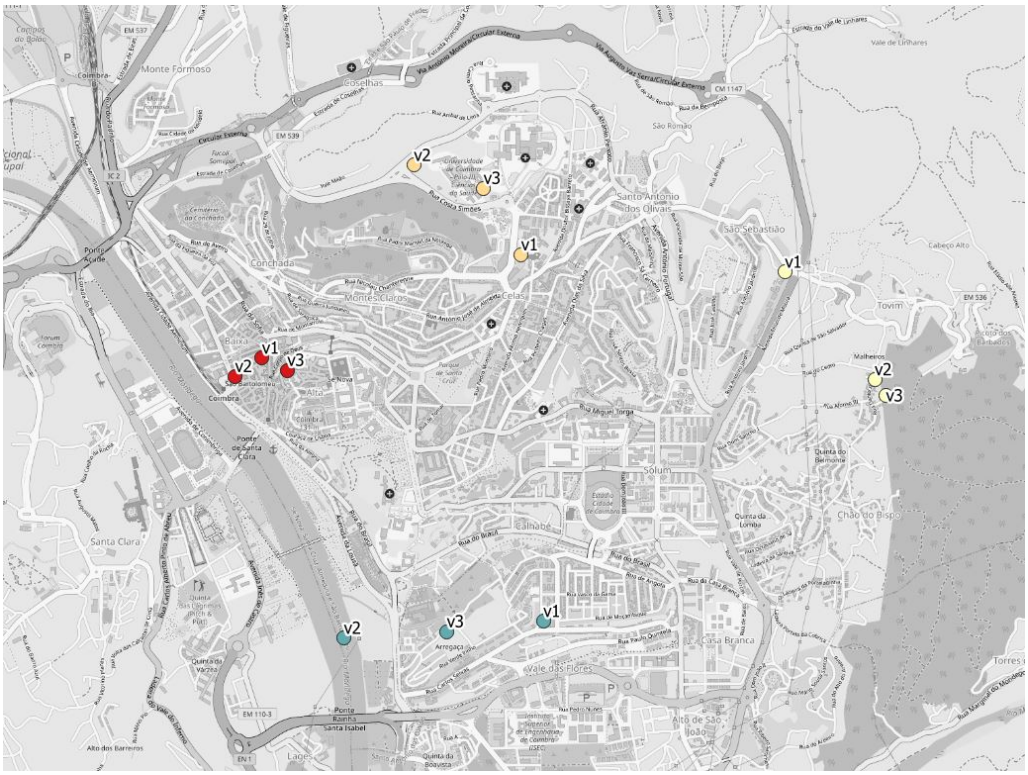


Figure 6.11: Detected home locations for 4 users using different positioning versions

Like previously described and exemplified in figure 5.7, in the original version, antennas are always assigned the tower's coordinates. In version 2, we change the coordinates of individual antenna's, with the centroid of the antennas signal area used as the new point. Finally in version 3, we used the centroid of the signal area (same as version 2) and the tower position to calculate the middle point between them. Figure 6.11 displays in a map layer the detected home locations, using each of the position versions. Four random users with homes in the city were selected for demonstration and are distinguished by the color of their respective circles. We can see by figure 6.11 that significant differences in home position occur when applying both versions 2 and 3 of antenna's positions. This difference would be incremented further outside of the city's urban area, where the average antenna radius is larger.

Table 6.9: Home validation for location versions

	Acc. Urban	Acc. Rural	Mean Distance	Std.	Q1	Q2	Q3
Version 1	66.55%	31.25%	4.509	8.747	0.600	1.384	3.432
Version 2	69.59%	52.96%	2.415	3.307	0.551	1.270	2.837
Version 3	75.12%	55.32%	2.230	3.340	0.439	1.022	2.411

Table 6.9 contains the results for running the validation algorithm for each of the three methods with Eps = 0.05, MinPts = 2 and 50km outlier filter. Both version 2 and 3 increase overall accuracy with bigger improvements noticed in rural areas by transitioning to individual antennas. This was to be expected in theory as areas with sparser coverage also have larger radius antennas. For these cases, the simple shift in position could sometimes be enough to bring the distance to a value that is within the success parameters. There is additionally a reduction of all distance related measurements, giving proof that the approach gives a good improvement when applied over the original CDRs.

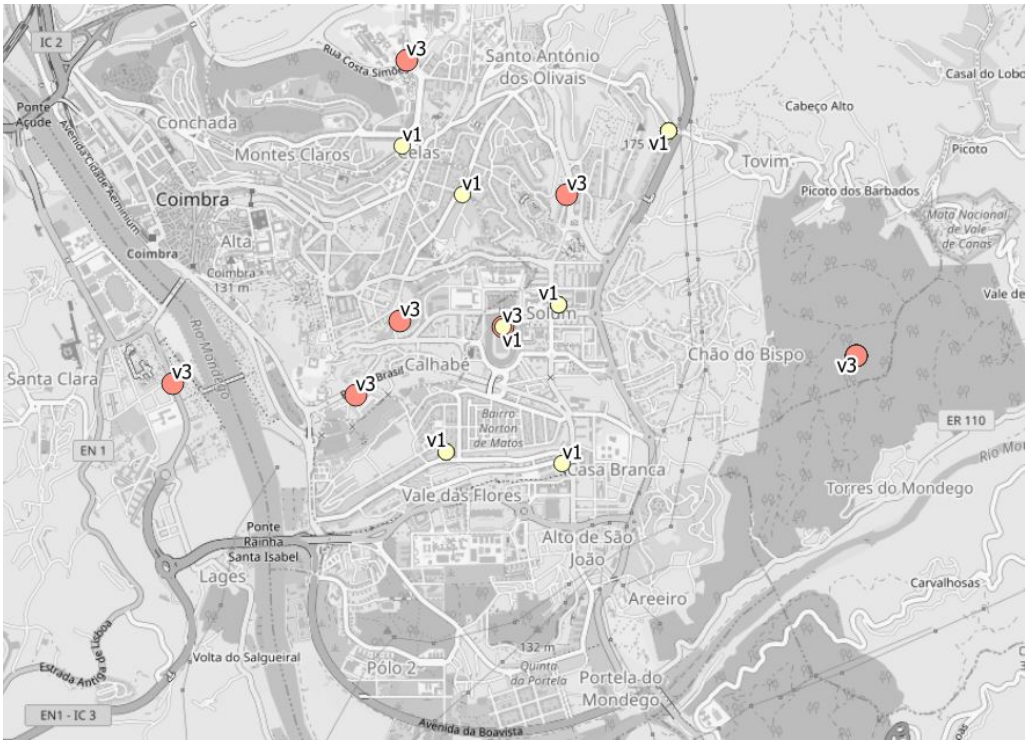


Figure 6.12: Detected Meaningful Places for one user using different positioning versions

There was additionally a comparison between versions for the detection of other routine places, an example of which is in figure 6.12. For one user, MVPs and OVPs are plotted indiscriminately with the original records position, version 1($v1$) and the individual antenna's position. In this case we used version 3 ($v3$) as it has obtained the best results in the first home validation algorithm. Although validation data for other routine places was not attained in time to be included in this work, the enhanced results with the home location give fundament that this method for improving CDRs could positively influence the accuracy of detecting other places as well.

This section, although not planed, became one of the major improvements and main differentiation factors from state-of-the-art methods. Its insertion in the overall methodology comes as a way to improve all other algorithms that use the coordinates present in Call Detail Records.

This page intentionally left blank.

Chapter 7

Conclusion

In this chapter, we summarize and reflect on the research made, reviewing and reiterating the key points of this work. First, we recapitulate the work developed and results obtained. Following this, we emphasize what our research has contributed to knowledge in this field. A section mentioning the challenges encountered and how we managed to surpass them is also present. Finally, we discuss some ways by which the work can be improved in the future.

7.1 Project Development

In this thesis, we addressed the objective of identifying and classifying places of interest using data from anonymised mobile communications events. We have presented a group of interacting methods, developing a system to analyse individual mobility behaviour. Such analysis relies on the locations present in Call Detail Records (CDRs) and intends to improve the understanding of user's regular places. Throughout the work we focused on several main steps, data pre-processing, the detection of home and workplace as well as other routine places, and finally, area classification.

Comparing the proposed and final Gantt diagrams (figure A.3 and figure A.4), we can see some discrepancies. We had to return to the optimisation and refinement of methods implemented in the first semester to adapt them to use the new reference tables. Furthermore, we needed also to develop an improved version of the location algorithm. The search and review of state-of-the-art methods took longer than expected, pushing back the final report. There was also a delay in model validation as we had not received this data until the end of May

Starting with data preparation, relevant methods for preparing records were applied, including a state-of-the-art method for detecting and excluding the Load Sharing effect. Even though the number of cases was not as large as initially expected, we succeeded in identifying these particular anomalies, as shown in section 6.1.1.

For the detection of home and workplace, we combine the criteria of defining a time interval of search [36] and Density Based Spatial Clustering or DBSCAN as used by the works [21] and [20] to group locations, described in section 5.1.3. A validation dataset allowed us to compare our method results with ground-truth data. Hence, our method appears to be on par or slightly above other relevant works with the same objective, as seen in section 6.2.1. The following step consisted of detecting the remaining routine places. As to not lose spatial precision, we chose an approach that did not consist of spatial clustering. Instead, we calculated a relevance metric for each location in the user's CDRs and grouped them into three separate categories inspired by others. This process had multiple iterations to find a suitable metric and output for classification but, we did not have access to further ground-truth data. For that reason, a comparative analysis between methods and validation of results could not be possible. Nonetheless, we are confident with the system developed, mainly when CDR location improvements are applied as these showed to improve the detection of home location.

The classification of areas by activity had the largest time allocation of any other phase of the project. This included deciding which POI data provider to use and completing the POI tables with opening hours. The proposed methodology for area selection suffered three main changes, each improving on the previous as flaws were encountered. The fixed radius approach was the easiest one to implement, however it did not consider that antennas have different coverage areas. The Voronoi diagram approach gave a hypothetical coverage area dependent on the density of antennas, but was fundamentally flawed in assuming users were closer to the antenna of connection than to any other. Finally, the circular sectors approach was implemented with data provided by the telecommunication service provider to create a good approximation of the signal area for each antenna. All three were conceptually compared in section 6.1.4. Once areas were defined, we extracted the POI related information, for each combination of time interval and day type, calculating a metric to achieve an activity label.

Every part of this work finally comes together in section 6.1.5, where we take user's routine places (both MVPs and OVPs) and classify those frequent places according to the time and day of visit. Ideally we would want annotated ground-truth data to compare these results of individual user mobility. This would allow us to optimise and change methods as done in the home and workplace detection.

7.2 Main Contributions

At an initial stage of our work, on the review of the State of the Art, we got familiar with the main approaches used to detect mobility patterns from CDRs and classify areas according to information in the contained POIs. This was essential to understand the challenges faced by others and potentially explore parts that could be improved in those works. Comparing this work with previous studies in the field, this one has several particularities not found in others.

For one, we had access to a high volume of data for both analysis and validation. This is something that allowed us to create methods that supported the high volume of data, retaining the performance if scalability is necessary. It also allowed us to make good claims in regards to the accuracy of our home and workplace detection method.

The circular sector approach, to create signal areas relative to each antenna is, according to our research, an innovative way to subdivide space for classification. It allows, in our understanding, for an improved match of the area where a user is when an mobile telecommunication event is made.

Additionally in this thesis we explored innovative improvements to the locations present in CDRs. Obtaining the approximate antenna signal area created the opportunity to further use this information to enhance user positioning in connection to antennas. This came to benefit already implemented methods and could possibly impact other future studies, giving better precision when applied over new CDRs.

7.3 Challenges

Due to the fact that data collection by the telecommunication service provider started in September, one challenge faced during this work was that the data provided originated from a time period during a worldwide pandemic situation. Even though the months of September/October of 2020 had a higher mobility than the stricter confinement period that followed, we know that they do not possess an accurate representation of pre-pandemic mobility.

Finally, a significant challenge that affects many studies attempting to infer mobility patterns from Call Detail Records is the attainment of ground-truth data to validate the obtained results. After requesting for various datasets that could validate our methods, we were able to get a home location dataset from the telecommunication service provider, whose amount and content was more that had previously been seen in other works. However, we are still left with several methods to validate, including routine locations and user's activities.

7.4 Future Work

It is important to reiterate some assumptions made in this thesis. For once, we assume during the period of study that we are not dealing with phones shared by more than one user and that no particular user changes or has more than one SIM card. Furthermore, there is the possibility that users, by any reason, did not make or receive calls in their workplace or home, precluding these places correct detection. The pandemic situation creates further possibilities that users worked mostly from home during the period of data collection.

Some possible improvements were found by conducting the analysis of area's activity classification. Manually giving a weight to POIs of certain types to increase their importance depending on the time of day, like restaurants at regular meal hours, possibly would change the activities to better mirror population tendencies. The same effect could also be achieved with a popularity/check-in value that was hour dependent, but as far as we know no POI dataset contains this information. There is still the question of points missing from the used dataset, as they might not be registered in the used data provider. One possible solution would be the combination of several POI datasets with the added difficulty of merging completely different category hierarchies into one.

The area's created for classification, although closer to the reality of where the user might be, still remain too large to have a good percentage of certainty in terms of user activity. Newer information sources that have been discussed with the telecommunication service provider for future work have the potential to improve the user's location even further. Doing so allows for a smaller search area and in general more accurate methods. The arrival of 5G networks, with more precise smaller radius antennas could be the next evolution step in mobility analysis using Call Detail Records. All methods and algorithms created and implemented have the foresight of easy adaptation for future technology's allowing continuation work to be carried out.

Bibliography

[1] Marta C. Gonzalez, Cesar Hidalgo, and Albert-Laszlo Barabasi. "Understanding Individual Human Mobility Patterns". In: *Nature* 453 (July 2008), pp. 779–82. DOI: 10.1038/nature06958.

[2] Deval Dixit. "Cross-layer design in cellular networks: issues and possible solutions". In: Dec. 2019.

[3] M. Tiru. "Overview Of The Sources And Challenges Of Mobile Positioning Data For Statistics". In: 2014, pp. 1–26.

[4] Francesco Calabrese, Laura Ferrari, and Vincent D. Blondel. "Urban Sensing Using Mobile Phone Network Data: A Survey of Research". In: *ACM Comput. Surv.* 47.2 (Nov. 2014). ISSN: 0360-0300. DOI: 10.1145/2655691. URL: https://doi.org/10.1145/2655691.

[5] A. K. Jain, M. N. Murty, and P. J. Flynn. "Data Clustering: A Review". In: *ACM Comput. Surv.* 31.3 (Sept. 1999), pp. 264–323. ISSN: 0360-0300. DOI: 10.1145/331499.331504. URL: https://doi.org/10.1145/331499.331504.

[6] Jiawei Han, Micheline Kamber, and Jian Pei. "10 - Cluster Analysis: Basic Concepts and Methods". In: *Data Mining (Third Edition)*. Ed. by Jiawei Han, Micheline Kamber, and Jian Pei. Third Edition. The Morgan Kaufmann Series in Data Management Systems. Boston: Morgan Kaufmann, 2012, pp. 443–495. ISBN: 978-0-12-381479-1. DOI: https://doi.org/10.1016/B978-0-12-381479-1.00010-1. URL: https://www.sciencedirect.com/science/article/pii/B9780123814791000101.

[7] Anil K. Jain. "Data Clustering: 50 Years Beyond K-means". In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by Walter Daelemans, Bart Goethals, and Katharina Morik. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 3–4. ISBN: 978-3-540-87479-9.

[8] Martin Ester et al. "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. Portland, Oregon: AAAI Press, 1996, pp. 226–231.

[9] Seongjin Park et al. "Quantitative Super-Resolution Imaging of Small RNAs in Bacterial Cells". In: vol. 1737. Feb. 2018, pp. 199–212. ISBN: 978-1-4939-7633-1. DOI: 10.1007/978-1-4939-7634-8_12.

[10] Hongzhi Shi et al. "Semantics-Aware Hidden Markov Model for Human Mobility". In: *IEEE Transactions on Knowledge and Data Engineering* 33.3 (2021), pp. 1183–1194. DOI: 10.1109/TKDE.2019.2937296.

[11] Christine Parent et al. “Semantic Trajectories Modeling and Analysis”. In: *ACM Comput. Surv.* 45.4 (Aug. 2013). ISSN: 0360-0300. DOI: 10.1145/2501654.2501656. URL: https://doi.org/10.1145/2501654.2501656.

[12] Desheng Zhang et al. “Exploring Human Mobility with Multi-Source Data at Extremely Large Metropolitan Scales”. In: *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking*. MobiCom ’14. Maui, Hawaii, USA: Association for Computing Machinery, 2014, pp. 201–212. ISBN: 9781450327831. DOI: 10.1145/2639108.2639116. URL: https://doi.org/10.1145/2639108.2639116.

[13] Gyan Ranjan et al. “Are Call Detail Records Biased for Sampling Human Mobility?” In: *SIGMOBILE Mob. Comput. Commun. Rev.* 16.3 (Dec. 2012), pp. 33–44. ISSN: 1559-1662. DOI: 10.1145/2412096.2412101. URL: https://doi.org/10.1145/2412096.2412101.

[14] Sahar Hoteit et al. “Filling the Gaps: On the Completion of Sparse Call Detail Records for Mobility Analysis”. In: *Proceedings of the Eleventh ACM Workshop on Challenged Networks*. CHANTS ’16. New York City, New York: Association for Computing Machinery, 2016, pp. 45–50. ISBN: 9781450342568. DOI: 10.1145/2979683.2979685. URL: https://doi.org/10.1145/2979683.2979685.

[15] Ghazaleh Khodabandelou et al. “Population estimation from mobile network traffic metadata”. In: *2016 IEEE 17th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM)*. 2016, pp. 1–9. DOI: 10.1109/WoWMoM.2016.7523554.

[16] Raul Montoliu and Daniel Gatica-Perez. “Discovering Human Places of Interest from Multimodal Mobile Phone Data”. In: *Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia*. MUM ’10. Limassol, Cyprus: Association for Computing Machinery, 2010. ISBN: 9781450304245. DOI: 10.1145/1899475.1899487. URL: https://doi.org/10.1145/1899475.1899487.

[17] Jun Zhang, Chun-yuen Teng, and Yan Qu. “Understanding User Spatial Behaviors for Location-Based Recommendations”. In: *Proceedings of the 22nd International Conference on World Wide Web*. WWW ’13 Companion. Rio de Janeiro, Brazil: Association for Computing Machinery, 2013, pp. 989–992. ISBN: 9781450320382. DOI: 10.1145/2487788.2488096. URL: https://doi.org/10.1145/2487788.2488096.

[18] Jiaxin Ding, Chien-Chun Ni, and Jie Gao. “Fighting Statistical Re-Identification in Human Trajectory Publication”. In: *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. SIGSPATIAL ’17. Redondo Beach, CA, USA: Association for Computing Machinery, 2017. ISBN: 9781450354905. DOI: 10.1145/3139958.3140045. URL: https://doi.org/10.1145/3139958.3140045.

[19] Denys Proux and Frederic Roulland. "Mobile Recommendation Challenges within a Strong Privacy Oriented Paradigm". In: *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Recommendations, Geosocial Networks and Geoadvertising*. LocalRec '19. Chicago, Illinois: Association for Computing Machinery, 2019. ISBN: 9781450369633. DOI: 10.1145/3356994.3365506. URL: https://doi.org/10.1145/3356994.3365506.

[20] Buddhi Ayesha et al. "User Localization Based on Call Detail Record". In: *Intelligent Data Engineering and Automated Learning – IDEAL 2019*. Ed. by Hujun Yin et al. Cham: Springer International Publishing, 2019, pp. 411–423. ISBN: 978-3-030-33607-3.

[21] Peiyu Yang et al. "Identifying Significant Places Using Multi-Day Call Detail Records". In: *Proceedings of the 2014 IEEE 26th International Conference on Tools with Artificial Intelligence*. ICTAI '14. USA: IEEE Computer Society, 2014, pp. 360–366. ISBN: 9781479965724. DOI: 10.1109/ICTAI.2014.61. URL: https://doi.org/10.1109/ICTAI.2014.61.

[22] Ana-Maria Olteanu Raimond et al. "Moving and Calling: Mobile Phone Data Quality Measurements and Spatiotemporal Uncertainty in Human Mobility Studies". In: vol. 2013. May 2013. ISBN: 978-3-319-00614-7. DOI: 10.1007/978-3-319-00615-4_14,.

[23] Maria Luisa Damiani et al. "On Location Relevance and Diversity in Human Mobility Data". In: *ACM Trans. Spatial Algorithms Syst.* 7.2 (Oct. 2020). ISSN: 2374-0353. DOI: 10.1145/3423404. URL: https://doi.org/10.1145/3423404.

[24] Daniel Orellana et al. "Uncovering Interaction Patterns in Mobile Outdoor Gaming". In: *2009 International Conference on Advanced Geographic Information Systems Web Services*. 2009, pp. 177–182. DOI: 10.1109/GEOWS.2009.13.

[25] Yu Zheng et al. "Recommending Friends and Locations Based on Individual Location History". In: *ACM Trans. Web* 5.1 (Feb. 2011). ISSN: 1559-1131. DOI: 10.1145/1921591.1921596. URL: https://doi.org/10.1145/1921591.1921596.

[26] Sibren Isaacman et al. "Identifying Important Places in People's Lives from Cellular Network Data". In: June 2011, pp. 133–151. ISBN: 978-3-642-21725-8. DOI: 10.1007/978-3-642-21726-5_9.

[27] Christian Quadri et al. "On Non-Routine Places in Urban Human Mobility". In: Oct. 2018, pp. 584–593. DOI: 10.1109/DSAA.2018.00075.

[28] Artjom Lind, Amnir Hadachi, and Oleg Batrashev. "A new approach for mobile positioning using the CDR data of cellular networks". In: *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*. 2017, pp. 315–320. DOI: 10.1109/MTITS.2017.8005687.

[29] Qihang Gu et al. “Inferring Venue Visits from GPS Trajectories”. In: *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. SIGSPATIAL ’17. Redondo Beach, CA, USA: Association for Computing Machinery, 2017. ISBN: 9781450354905. DOI: 10.1145/3139958.3140034. URL: https://doi.org/10.1145/3139958.3140034.

[30] Santi Phithakkitnukoon et al. “Activity-Aware Map: Identifying Human Daily Activity Pattern Using Mobile Phone Data”. In: vol. 6219. Aug. 2010, pp. 14–25. ISBN: 978-3-642-14714-2. DOI: 10.1007/978-3-642-14715-9_3.

[31] Yihong Wang et al. “Understanding travellers’ preferences for different types of trip destination based on mobile internet usage data”. In: *Transportation Research Part C Emerging Technologies* 90 (May 2018). DOI: 10.1016/j.trc.2018.03.009.

[32] Yang Xu et al. “How Friends Share Urban Space: An Exploratory Spatiotemporal Analysis Using Mobile Phone Data”. In: *Transactions in GIS* 21 (June 2017). DOI: 10.1111/tgis.12285.

[33] Xingang Zhou et al. “The Uncertain Geographic Context Problem in Identifying Activity Centers Using Mobile Phone Positioning Data and Point of Interest Data”. In: June 2015, pp. 107–119. ISBN: 978-3-319-19949-8. DOI: 10.1007/978-3-319-19950-4_7.

[34] Guang Yuan et al. “Recognition of Functional Areas Based on Call Detail Records and Point of Interest Data”. In: *Journal of Advanced Transportation* 2020 (Apr. 2020), pp. 1–16. DOI: 10.1155/2020/8956910.

[35] Mi Diao et al. “Inferring individual daily activities from mobile phone traces: A Boston example”. In: *Environment and Planning B: Planning and Design* 43 (Sept. 2015). DOI: 10.1177/0265813515600896.

[36] Maarten Vanhoof et al. “Detecting home locations from CDR data: introducing spatial uncertainty to the state-of-the-art”. In: (Aug. 2018).

[37] Marco Mamei, Massimo Colonna, and Marco Galassi. “Automatic Identification of Relevant Places from Cellular Network Data”. In: *Pervasive Mob. Comput.* 31.C (Sept. 2016), pp. 147–158. ISSN: 1574-1192. DOI: 10.1016/j.pmcj.2016.01.009. URL: https://doi.org/10.1016/j.pmcj.2016.01.009.

[38] Mehdi Khosrow-Pour D.B.A., ed. *Encyclopedia of Information Science and Technology, Third Edition:* IGI Global, 2015. ISBN: 9781466658882 9781466658899. DOI: 10.4018/978-1-4666-5888-2. URL: http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-4666-5888-2 (visited on 06/10/2021).

[39] Renato Andrade, Ana Alves, and Carlos Bento. “POI Mining for Land Use Classification: A Case Study”. In: *ISPRS International Journal of Geo-Information* 9.9 (2020). ISSN: 2220-9964. DOI: 10.3390/ijgi9090493. URL: https://www.mdpi.com/2220-9964/9/9/493.

[40] Filippo Maria Bianchi et al. "Identifying user habits through data mining on call data records". In: *Engineering Applications of Artificial Intelligence* 54 (2016), pp. 49–61. ISSN: 0952-1976. DOI: https://doi.org/10.1016/j.engappai.2016.05.007. URL: https://www.sciencedirect.com/science/article/pii/S0952197616300975.

[41] URL: https://www.usna.edu/Users/oceano/pguth/md_help/html/approx_equivalents.htm.

[42] URL: https://www.quora.com/How-many-meters-make-up-a-degree-of-longitude-latitude-on-Earth.

This page intentionally left blank.

Appendix A

Gantt Charts

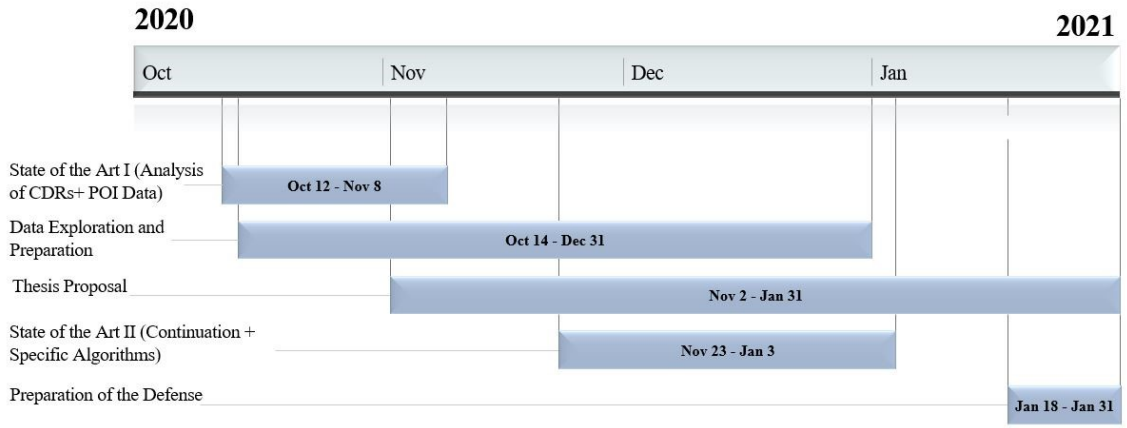


Figure A.1: Proposed Gantt chart of the first semester.

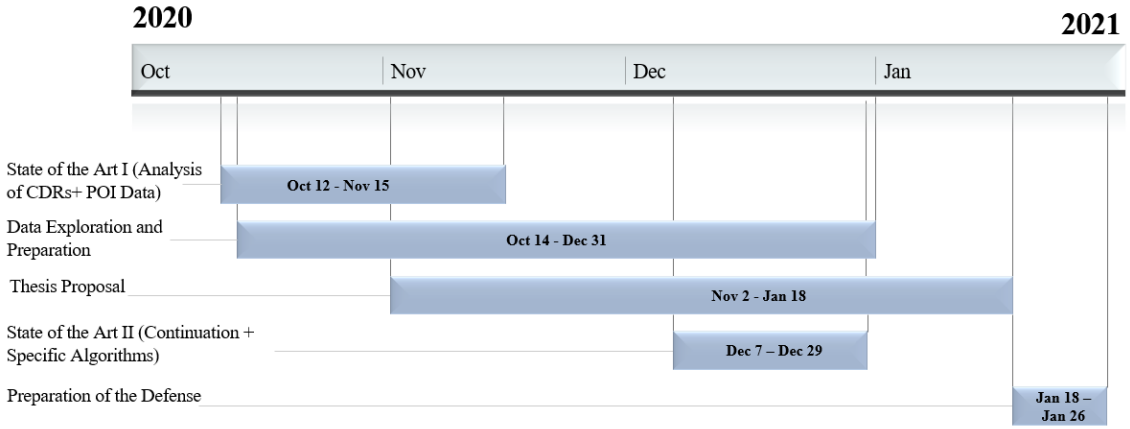


Figure A.2: Final Gantt chart of the first semester.

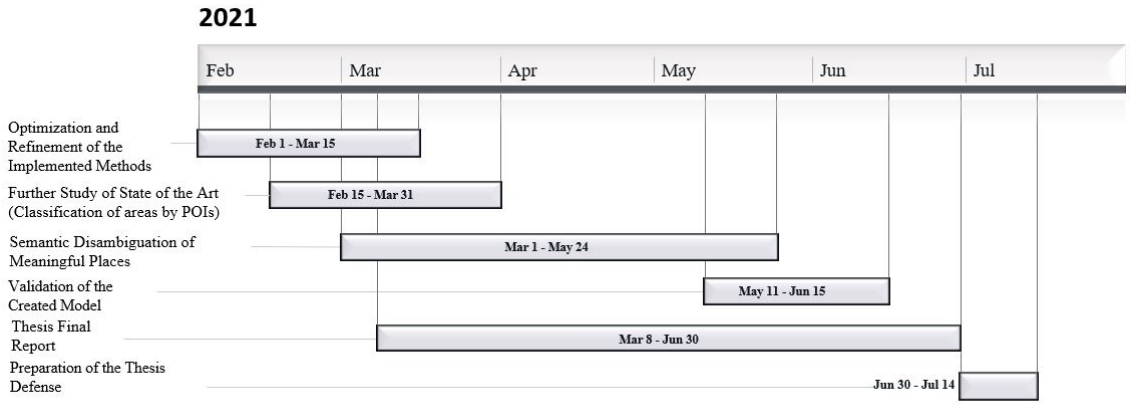


Figure A.3: Proposed Gantt chart of the second semester.

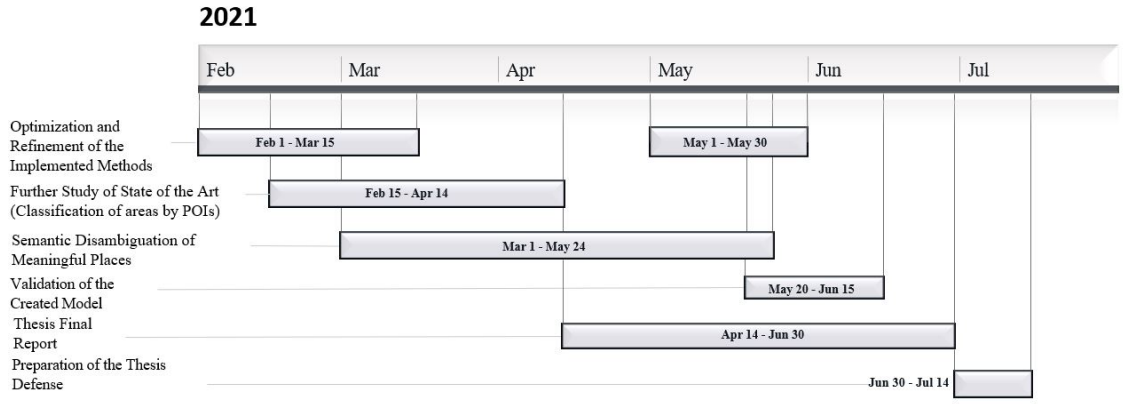


Figure A.4: Final Gantt chart of the second semester.