UNIVERSIDADE D
COIMBRA

Cláudia Beatriz Almeida Rodrigues

My Places

Identification of User's Geographic Map

Dissertation in the context of the Master in Informatics Engineering, Specialization in Communications, Services and Infrastructures advised by Professor Doctor Carlos Bento and Professor Doctor Marco Veloso and presented to the Faculty of Sciences and Technology Department of Informatics Engineering

June 2021

This page is intentionally left blank.

Abstract

Cities are becoming more and more centered on sustainability, with the additional goal of making them creative and vivid places for their citizens and visitors. Having this in mind, it is important to study how citizens use and move on urban spaces. The capacity of improving these spaces relies, overall, on having data that can support decision-making and provide valuable information to plan a city. The data can also provide insights to private companies in multiple ways, especially on how to run their businesses.

Telecommunication companies have been struggling to find ways to attract clients. A strategy that they adopt is the study of their clients to customize services. This is frequently made through the analysis of ubiquitous data, including Call Detail Records (CDRs), that they usually collect for billing purposes. Some companies use CDRs to create and analyze new types of data sources, like Snapshots, that attempt to deal with the sparsity and irregularity of the mentioned data.

Geospatial data, such as CDRs, plays a crucial role in studying land use and urban flows, and even in comprehending behaviors, habits, preferences, and needs of individuals. The processing and analysis of this data can be used to infer places where people spend most of their daily time and to get an overview of human trajectories and crowd movements.

The fact that the telecommunication companies have started cooperating more and sharing some of their data is a win-win situation for them, cities, and the researchers that develop this type of work.

Throughout this work, we study the usage of CDRs and Snapshots to identify important places in individuals' life or, in other words, places that are regularly visited by them. We use these data sources, provided by a telecommunication company, to identify their subscribers' geographic map (geo-profile) by identifying home, second home, and work zones, at an antenna level. The geo-profile is accomplished after determining the profile of the user (day worker, night worker, etc.) according to his/her activity in the network.

Outcomes of this thesis are the segmentation of customers, to profile them based on their daily CDRs and Snapshots, and a model, that uses clustering, to identify their geo-profile also based on these data. Lastly, we use ground-truth to validate and evaluate the results on the inference of home locations, and we reached an accuracy of 66% with the methodology applied.

Keywords

Call Detail Records, Clustering, Data Analysis, Geo-Data, Geo-Profile, Home, Land Use, Meaningful Places, Mobile Data, Second Home, Sleeping Period, Workplace.

This page is intentionally left blank.

Resumo

As cidades estão a tornar-se cada vez mais centradas na sustentabilidade, com o objetivo adicional de as converter em locais criativos e vívidos para os seus cidadãos e visitantes. Tendo isso em mente, é importante estudar como os cidadãos usam e se movimentam nos espaços urbanos. A capacidade de melhorar estes espaços depende, geralmente, da disponibilidade de dados que suportem a tomada de decisões e forneçam informações valiosas para o planeamento de uma cidade. Os dados também podem fornecer informações a empresas privadas de várias formas, especialmente sobre como administrar os seus negócios.

As empresas de telecomunicações têm lutado para encontrar maneiras de atrair clientes. Uma estratégia que adotam é o estudo dos seus clientes para customizar serviços. Isto é frequentemente feito através da análise de dados ubíquos, incluindo Registos de Detalhes de Chamadas (CDRs), que eles geralmente recolhem para fins de faturamento. Algumas empresas usam CDRs para criar e analisar novos tipos de fontes de dados, como os Snapshots, que tentam lidar com a dispersão e irregularidade dos dados mencionados.

Dados geoespaciais, como CDRs, desempenham um papel crucial no estudo do uso do solo e fluxos urbanos, e até mesmo na compreensão de comportamentos, hábitos, preferências e necessidades dos indivíduos. O processamento e a análise destes dados podem ser usados para inferir lugares onde as pessoas passam a maior parte do seu tempo diário e até mesmo para obter uma visão geral das trajetórias humanas e dos movimentos da população.

O facto de as empresas de telecomunicações terem começado a cooperar mais e a partilhar alguns dos seus dados, é uma situação vantajosa para elas, as cidades e investigadores que desenvolvem este tipo de trabalho.

Ao longo deste trabalho, estudamos o uso de CDRs e Snapshots para identificar lugares importantes na vida dos indivíduos ou, por outras palavras, lugares que são regularmente visitados por eles. Utilizamos estas fontes de dados, fornecidas por uma empresa de telecomunicações, para identificar o mapa geográfico dos seus assinantes (geo-perfil), identificando zonas de casa, segunda casa e de trabalho, a nível de antena. Este geo-perfil é realizado após a determinação do perfil do indivíduo (trabalhador diurno, trabalhador noturno, etc.) de acordo com sua atividade na rede.

Os resultados desta tese são a segmentação dos clientes, para traçar o seu perfil com base nos seus CDRs e Snapshots diários e um modelo, que usa agrupamento, para identificar seu geo-perfil também com base nesses dados. Por fim, usamos dados verdadeiros para validar e avaliar os resultados da inferência da localização de casa, onde alcançámos uma precisão de 66% com a metodologia aplicada.

Palavras-Chave

Registos de Detalhes de Chamadas, Agrupamento, Análise de Dados, Geo-Dados, Geo-Perfil, Casa, Ocupação Territorial, Sítios Importantes, Dados Móveis, Segunda Casa, Período de Descanso, Local de Trabalho.

This page is intentionally left blank.

Acknowledgements

I would like to thank my advisor, Professor Carlos Bento and my co-adviser, Professor Marco Veloso for their patience, dedication, and their valuable guidance throughout this project. I would also like to thank Professor Ana Alves for her availability and dedication.

I would like to send a huge thanks to my colleagues from the Department for the support and the afternoons that we spent playing cards. Also, a huge thanks to Rodrigo that never let me quit, and to Carolina, Catarina, Nicole, and Bruna for texting me every day of this journey just to remind me that life is amazing.

Finally, I would like to express my sincere gratitude to my family, especially to my parents who have always supported me and push me to go further. Ultimately they are the reason behind my conquers.

This page is intentionally left blank.

Contents

Acronyms

AmILab Ambient Intelligence Laboratory. 3, 49, 50, 52

CDRs Call Detail Records. iii, xv, xvii, 1–13, 23, 26–29, 45, 49–53

CISUC Centre for Informatics and Systems of the University of Coimbra. 3

CLARANS Clustering Large Applications based on RANdomized Search. 16

csv Comma-Separated Values. 23, 44

CURE Clustering Using REpresentatives. 16

DBMs Database Management Systems. 17

DBSCAN Density Based Spatial Clustering of Applications with Noise. 10, 12, 14, 17–21, 33, 34, 36, 45, 47, 49

DENCLUE Density-based clustering. 20

eps Epsilon. 10, 17–19, 21, 22, 32–34, 42, 47

GIS Geographic Information System. 37

GPS Global Positioning System. 1, 5, 6, 12

GSM Global System for Mobile Communications. 1, 9

HDAs Home Detection Algorithms. 10, 11

OPTICS Ordering Points To Identify the Clustering Structure. 20

SSE Sum of Squared Error. 15

STING Statistical Information Grid Approach. 20, 21

VDBSCAN Varied Density Based Spatial Clustering of Applications with Noise. 14, 18, 19, 21, 22, 32–34, 42, 47, 49, 50

WCSS Within Cluster Sum Of Squares. 32

This page is intentionally left blank.

Glossary

algorithm A logical arithmetical or computational procedure that if correctly applied ensures the solution of a problem. 4, 5, 9–22, 29–34, 36, 39, 43, 47–49

method Procedure, usually according to a definite, established, logical, or systematic plan. 2, 7, 8, 10, 12, 13, 16, 18, 22, 31–33, 35, 36, 40, 49–52

methodology The system of methods and principles used in a particular discipline. iii, 2, 3, 31, 39, 50

model Informative representation of an object, person or system. iii, 1–3, 5, 8–10, 12, 14, 22, 23, 31, 32, 42, 44, 47–53

technique A practical method, skill, or art applied to a particular task. 2, 5, 7, 9, 12, 14, 16, 21, 29, 31, 32, 34, 50, 51

This page is intentionally left blank.

List of Figures

This page is intentionally left blank.

List of Tables

This page is intentionally left blank.

Chapter 1

Introduction

The trajectories in the human routines tend to be quite predictable, presenting a high degree of temporal and spatial regularity [14] [24]. The study of this trajectories, generally leads to the creation of models to identify patterns taken by individuals that, most of the time, can faithfully reproduce their movements. Ultimately, this is a task that relies on geospatial data, sometimes obtained from Global Positioning System (GPS) devices or from Global System for Mobile Communications (GSM). Call Detail Records (CDRs) are a type of data that is collected by mobile phone operators for billing purposes and used in location analytics to characterize various aspects of human mobility [31].

The examination of human mobility from mobile phone usage can be applied in areas as varied as urban planning, ecology, or mobile sensing. It also addresses daily individuals activities and helps to improve sustainability and to build smart cities [17].

The present document describes a process for individual geo-profiling, using the users' records on the mobile network. The chapters and sections of this document describe in detail the work under the context of this thesis.

This chapter gives the reader a brief introduction to the project and explains the main lines for this research. It starts by presenting the motivation behind this study, the objectives that were established, and the outcomes expected. Then, the course and context of the internship are briefly described. The last section describes the structure of this document.

1.1 Motivation

The regularity in human movements, shows that people spend most of their time in a small number of places, mostly home and work. Specially in urban places, the distribution of meaningful places and the movement between them can be aggregated to identify the spatial structure and activity patterns in cities. This information is very useful in transport and urban planning, environmental policies, and business advertisement [34].

The extraction of places that are regularly visited by users, can be the basis to understand the mobility patterns in the area. The idea of this project is to geo-profile individuals by identifying their meaningful places, such as home, second home, and workplace. Moreover, the identification of these places, especially home and work, is an important step to develop mobility patterns and infer recurrent commuting trips of the individuals [25]. This motivation also comes from the competition that telecommunication companies are facing in the marketing environment.

Since the offer of products has increased in the telecommunication sector, companies can no longer rely on the strategy of only attract new clients: sometimes, more important than increasing the market share, is the development of strategies focused on retaining their own customers and capture back the attention of former subscribers. By geo-profiling clients, telecommunication companies can recognize mobility characteristics and meaningful places, taking advantage of this information to improve and adapt marketing campaigns or quality of service. Further, this knowledge is going to give them the chance of offering new and more customized services, such as new phone plans or internet for a second home location, taking into account the client's needs or habits.

At a research level, some aspects motivated the development of this project, such as scientific challenges as the sparsity and irregularity of CDRs or the validation and evaluation of the model. Although some published articles aimed to identify human patterns and places that are important in people's daily lives through the analysis of CDR data, the truth is that researchers still have controversial opinions on using this type of data for this matter. Adding to that, some authors that used this data, did not validate their work mainly due to the lack of ground-truth data [39] [36], which increases even more the distinct opinions.

Another motivation is that, even though nowadays it is common for people to have more than one job or home location, most of the published works focused their attention on finding only one location for home and one location for work [39]. Furthermore, clients from different countries have different routines, so different companies are developing their own methods, by studying their own clients.

1.2 Objectives

The main goal of this work arises from the above motivations and is to identify the geographical map of a set of subscribers from a telecommunication company, in a way that it is possible to determine their individual profile, by determining if the client is a day or night worker based on the CDRs activity and, from that, build his/her geo-profile. To build the geo-profile, it is expected that we identify the important places of each user, which means the locations where he/she spends most of his/her time, including home, second home, and workplace without assuming that he/she has only one home or one job.

To achieve this goal, we will start by reviewing the approaches and the current state-of-the-art on detecting meaningful places in people's daily life using CDRs. After understanding the applied approaches, we pretend to find innovative ways and optimized methodologies to develop a model. Once the methodology is defined, the necessary methods and techniques, including data mining techniques, such as clustering, will be applied to infer important places in each client's life. Despite its application being made only to a set of subscribers, the model must be effective and scalable when applied to larger groups of subscribers. Another objective is to ensure that the results are trustworthy, so a phase of the project will be dedicated to the validation and evaluation of the outcomes.

In this project, we will infer about geographical characteristics of a set of users that have most of their CDRs registered in Coimbra, Portugal. The outcomes must provide the necessary information about each client most frequented locations, such as home, second home, and work locations, in a way that we can understand the user's dynamics and the telecommunication company knows which are the adequate services to offer.

1.3 Outcomes

The outcomes of this project aim to provide resources for various stakeholders. This work is integrated into an internship at Ambient Intelligence Laboratory (AmILab) and counts with the participation of a telecommunication company laboratory, so, both will use the resultant elements for decision-making in the respective area.

The main outcome of this project is a final model for geo-profile users. This model must identify significant places in their lives or places where they spend most of their time, like home and work. This identification is made at an antenna level, which means that we will not determine the exact location of the place, but the zone where is most probable that the place is located.

However, throughout the process, it is also expected the achievement of secondary outcomes, such as the determination of the users' sleeping period and their profile according to that. With the results of this work, the labs must be able to, given a set of CDRs belonging to multiple subscribers, determine each user's geographical map (geo-profile), by identifying their home, work, and if applicable, second home locations, through the characterization of their routine (day worker, night workers, etc.) according to phone activity.

The outcomes of the model have to be efficient, trustworthy, and scalable. This way, every time a user is selected to be geo-profiled, all of the mentioned information has to be returned.

To achieve these outcomes, the results of the model have to be validated and evaluated with ground-truth. Therefore, the validated results can help, for example, transport operators to adapt the offer, business related to land use, and even the telecommunication companies to customize their services.

In a scientific context, it is also expected that we write a scientific paper based on the work developed throughout this thesis.

1.4 Internship

This internship took place at AmILab, a research lab dedicated mainly to ambient intelligence, pervasive computing, and ubiquitous computing. This lab is part of the Centre for Informatics and Systems of the University of Coimbra (CISUC) and focuses its work on understanding and predicting relevant dynamics, especially in the urban areas, for urban planning, land use, and intelligent transport systems.

This project was planned and scheduled by the AmILab team. The team performed weekly meetings where all members discussed the progress achieved during the week, challenges, and possible future work. Every two weeks, we performed a meeting with the telecom company to report the state of the project, including progress on the work done, risks, and future work.

Furthermore, at the beginning of the project, we performed a design thinking session between the AmILab members and the telecom company members, in the format of a workshop, to consolidate and align thoughts and expectations on this project.

Along with the development, we also attend sessions with multiple stakeholders involved in this project and other projects, to present the achievements and the methodology applied.

1.5 Document Structure

This document is organized and divided into six chapters and is gives an overview on the work developed for this thesis throughout the internship. This first chapter gives a brief introduction to the structure and the chapters of this thesis. It introduces the project, describing its context, the motivations to geo-profile users, objectives, the outcomes established, and the course of the internship.

The second chapter presents the state-of-the-art. It starts by presenting an overview on data sources, including some of the authors that use CDRs to determine important places, and the controversial opinions on this data. It also presents the Snapshots, a new type of data provided by the telecom company. Then, it describes some approaches that other researchers used to determine meaningful places and a strategy that some companies follow to know their clients better, which consists of segmenting clients with the intuition of profiling them. The last part of this chapter is dedicated to spatial clustering and it briefly describes some clustering algorithms, given especially attention to the ones that were mentioned on the approaches for meaningful places and the segmentation. At the end of the chapter, it is presented a succinct comparison between the algorithms that were mentioned in the literature described.

The third chapter is dedicated to the description and analysis of the data that we have available. It is also in this chapter that we start presenting the processing of the data.

The fourth chapter describes the methodology, approach, and tools used to develop this thesis.

The fifth chapter presents the outcomes of the project, including the geo-profiling, validation and evaluation, and scalability results.

The sixth and last chapter presents the conclusions that were achieved, the main contributions of this work, some of the challenges that were faced, and a description of the possible future work.

Chapter 2

State-of-the-art

Discovering patterns from records of movements can be a challenging task that continues to be the basis or, at least, part of many academic research papers. This chapter addresses the state-of-the-art for this work, comprising an overview on the data sources that we have available, the methodologies applied by other authors to identify meaningful places (mostly home and work) including the algorithms used for spatial clustering, and an approach to segmentation clients.

2.1 Data Sources

Passively-generated data is usually collected for billing purposes and includes data that is collected from internet browsers, web sites, and mobile phones. Despite not being specifically generated for inference of mobility and social patterns, is commonly used for that and for studying human behavior and land use.

Given the mobile phones worldwide availability and their ubiquitous aspects, the study of human movements through the exploitation of mobile phone data has been an active area of research in the last few years [20]. Some techniques associated directly with the user's phone, for example, the GPS or location services, when active provide data that can be attractive due to the precision of traces. However, it compromises the users' privacy and requires their participation in the processes of giving consent and activating the service. On the other hand, CDRs can be used as an alternative. This data source is provided by mobile operators, without requiring the users' participation, with the advantages of being available for significant parts of the population and once anonymized is not intrusive.

Another type of data used to build models is the active-generated data. Although it does not contribute directly to identify important places or trajectories, its use is very important to validate and evaluate the results. In this type of data, the identity that is collecting the data "creates" the information that is provided, as the request of the information is made to obtain a specific type of answers. For example, ground-truth data can be considered active-generated data because most of the time is generated and collected specifically for validation and evaluation purposes.

2.1.1 Call Detail Records (CDRs)

Call Detail Records (CDRs) are a type of data containing non-continuous traces that is generated every time a mobile subscriber connects to an antenna. These records include immense information on how, when, where, and with whom people communicate daily. In other words, their analysis provide knowledge on the user's sent and received calls and text messages, as well as about the telephone mast that received/transmitted them. Given the sensitivity of the information that this data contains, it is a frequent and good practice the anonymization of the identifying fields to guarantee the confidentiality of the data and protect the subscriber's identity [32]. Depending on the identity that is extracting the data and the final purpose of it, the records can contain various columns with different information about calls and text messages. Meanwhile, there is some information that is essential to identify meaningful places: the id of the user, the date and hour of when the event was registered, and the location of the cell tower where the call was registered (latitude and longitude). Table 2.1 presents an example of a CDRs sample with the mentioned columns.

Table 2.1: CDRs Sample

USER ID	TIMESTAMP	CELL ID	LAT	LONG
91961FG	14/09/2020 12 : 03 : 23	568396	40.16045	-8.85134
81G8LO7	13/10/2020 21 : 33 : 28	675359	40.38648	-8.22967

Many researchers and institutions are aware of the potential and efficiency of CDRs and prefer this data containing spatial-temporal information to discover human mobility patterns rather than other types of data, mostly because it requires least investments [39]. As mentioned above, it can be used as an alternative to GPS, and besides both types of data (CDRs and GPS) being considered "data rich" and both having potential in researching human mobility patterns, CDRs have significant limitations as a source of location information [16] [32] [6]:

- Temporal irregularity and sparsity: The information is only generated when the user interacts with the mobile phone (the temporal regularity of the annotated information is not technological given, but a result of the user's behavior);
- Spatial sparseness: Only information on the cell tower where the call or the text message was registered is available. The coverage area of a cell varies substantially between different towers and connections modes, which makes it very hard to obtain exact locations;
- Non-routine events: Contrary to regular events, like going to work or home, non-regular events which are not part of the usual routine trajectories, are difficult to predict.

As a result of the uniqueness of its advantages, another challenge concerned with ground-truth is raised. The granularity in which these statistics are available makes it almost impossible to find matching data to validate the results.

As a matter of fact, there are controversial opinions about the use of CDRs to study human mobility. Although, a study made to investigate possible caveats and limitations of CDRs, revealed that the underlying nature of the data introduces a certain degree of bias, which probably occurs due to the uneven distribution of people's interaction with the network [41], the outcomes of this research were not considered conclusive. Despite these challenges,

some researchers and institutions ended up proving that it can reflect human mobility and significant places [31], considering this data representative and using it to achieve goals, especially when it comes to mobility and identification of meaningful places [16] [39] [34] [36].

Given that CDRs are routinely recorded, network-wide and provide direct access to localization data for large samples of the population, this data can be very attractive for large-scale analysis of individuals location [36], and not only allows the analysis of single entities but also communities on a large scale [22]. Considering this fact, some researchers focused their work on studying CDRs data to extract types of social communities in the network, their relationships, and movement patterns including places of residence (homes) and work places [37] [18].

To name specifically some contributions considering CDRs that are directly related to this project, there are the identification of home and workplaces to calculate the distribution of commute distances and estimate the carbon footprint of Los Angeles and New York [16], the extraction of user's trajectories and identification of significant places of Beijing's (China) population [39], and the exploration of home and work location of users from Bangkok (Thailand) [34]. Besides the mentioned contributions there are others that contributed directly to transport planning [5] or health care [33]. These works provided multiple results, because they include diverse empirical methods and frameworks to extract spatio-temporal insights from mobile data.

There are many possible fields that benefit from mobile phone network data. Overall, the study of individuals and communities are areas of big interest to develop smart cities with intelligent transports, ecologically sustainability, and opportunities to establish and expand business not only in cities but also in more rural areas.

2.1.2 Snapshots

Records of mobile phone usage are sometimes "bursty" depending on the number of records registered at certain hours of the day, with close temporal proximity, which can result in an over-representation of certain cell towers. To face this challenge, Furletti et al. [13], partitioned the day into equally sized intervals and every time slot had the capacity of representing only one CDR. Burkard et al. [6] followed the same reasoning and, for each time slot, chose the CDR register that was closest to the center of the slot, to represent it.

Similar to the previously mentioned authors, a telecom company adopted an identical technique to collect information, and called the data "Snapshots". A Snapshots dataset is composed of voice calls, SMSs, and connections or disconnections of the mobile phone from the wireless network. These different interactions are called events and are gathered every two hours, starting at midnight each day (meaning that every two hours a snapshot is registered). The information registered in each snapshot consists of the last event that was made over the two-hour slot. If no event is recorded during the two hours, it is registered the last event made by the client. After 12 hours without new events, the records of that client are no longer collected until a new event occurs. Table 2.2 presents a possible structure of the dataset, which is similar to the structure of a CDRs dataset, however, it has an additional column that represents the time when the snapshot was recorded, the "DUMPTS" (YYYYMMDDHH24).

Although the information provided by the snapshots is still being exploited, the telecom company that has been recording it, has been having promising results when using this type of data. Like CDRs, Snapshots come with the asset of being real data and most of

Table 2.2: Snapshots Sample

USER ID	TIMESTAMP	DUMPTS	CELL ID	LAT	LONG
91961FG	2020/09/03 21 : 32 : 20	202009040800	568396	40.16045	-8.85134
81G8LO7	2020/09/03 23 : 35 : 48	202009040800	675359	40.38648	-8.22967

the time updated, which adds comfort to the results obtained.

2.1.3 Ground-truth Data

Ground-truth data is often used to help building models, overall by validating results. As mentioned, CDRs are commonly used to capture people's information, but the validation of the results cannot be furnished from this type of data [8]. On the other hand, the detailed and precise information provided by, for example, data collected by companies containing information about their post-paid clients, makes it an attractive source to be employed as ground-truth, and with it perform the validation and evaluation, which are a crucial part when developing this type of projects.

Other data that can be used to perform the validation and evaluation is the surveys, which when available can be very valuable. There are at least two ways of using surveys to validate and evaluate results: some works perform the validation and evaluation using surveys that include the entire population of a country, which are named census [16] [36] and others use surveys with collected information from volunteers representing a sample of the population [16] [25].

2.2 Determining Meaningful Places

As mentioned previously, recognizing how and when people move, where they move, and where are their important places, such as home and work, can be crucial for business purposes, land use decisions, transport infrastructures, or the definition of urban policies. There are several approaches that were developed to describe and analyze people's daily movements and extracting from them home and workplaces.

In practice, the identification of an important place means that one or more cell towers are identified as a place/region where the user spends a considerable amount of time. Typically, the home or workplaces are considered the important places and their identification is based on the user's activities. This section describes some of the proposed approaches to identify important places, at an antenna level, using CDRs.

2.2.1 Data Processing

Before starting the development of models, the authors normally start by investigating and defining their methodology. Part of this process begins with the analysis of the available data, then it is usual that the data is treated according to what the methods require. This data processing task can take place throughout the entire project.

Depending on the analysis to be performed, it can be necessary to restrict the user base, by removing, from the study, clients that do not have enough activity or an elevated number of activities.

Some researchers considered that users with a low number of events can further represent noise to the results or users that have more than a certain number of activities are probably companies that are being treated as individuals, so they excluded them from the datasets.

Related to this subject, Ahas et al. [2] proposed a method to identify anchor points (e.g., Home and Work) and assumed that this was an impossible task to perform in users with lower activity and users that have to many calls, considering that the last ones represented an organized call procedure or a technical device using GSM network. So, they removed both of those correspondents, by eliminating the ones that had registers of events in their most visited cell on fewer than seven days a month and, users with more than 500 events per month. This was made based on the average number of events made per month by one person, which was 81.

Isaacman et al. [16] omitted from their dataset, the ID's of users that were registered as business numbers, retaining only phones registered to individuals. Then, removed registers of clients that appeared in their base ZIP code, less than half the days they had events, assuming that these users moved to other parts of the country, but retained their old billing addresses. After the filtering step, they considered their CDRs dataset "an useful representation of mobility and telephone usage in the regions of interest".

Zhao et al. [41] considered that if the footprints generated by a respondent in a day did not provide acceptable temporal coverage, his registers should be removed from the dataset. For that, they divided the day into six-hour slots, and only those subscribers with at least one event registered in each slot were included in the study.

Another work that take this filtering process, was Lumpsum et al. [34], by filtering users with an equal frequency to the minimum usage over 12 hour slots or whose minimum usage was more than half of the maximum usage.

2.2.2 Home and Workplaces

Depending on the objective that is established, after the analysis and pre-processing of the data, there are different paths that researchers take to build their models. Although CDRs may not sample all locations corresponding to user's visits, owing sparsity in time and space of voice calls and text messages, several techniques have been surging to more accurately discover human movements and locations. This section describes some contributions that used this data to identify home and work locations.

Sometimes, authors rely on the user's common behavior, using *a priori* assumptions (e.g., criteria with temporal constraints) to determine important places. A research work in this matter that adopts criteria with temporal constraints is from Isaacman et al. [16]. In an initial phase, they used data provided by 37 volunteers to propose and evaluate a model for important places analysis, then, to test their approach, they apply it on a more universal dataset containing CDRs of users from Los Angels and New York.

This work aggregated different algorithms. The first that was applied, identifies important places and is divided into two stages: (1) clustering cell towers that appear in the user's trace and (2) identifying the importance of clusters (based on a logistic regression over the volunteer's CDRs). After, they developed their own algorithms to apply semantic meaning to the important clusters (places) identified, namely Home and Work. The home detection algorithm chose the cluster with higher number of events during the time frame applied for "home hours", which was from 7PM to 7AM. The work detection algorithm first ranked all clusters based on the number of work-hour events that were made on weekdays from

1PM to 5PM, and then the cluster with higher number of events in the work-hours and lower number in the home-hours was assigned to be the workplace location.

Moreover, the model was validated and evaluated by comparing the results to US census and ground-truth data provided by the volunteers, demonstrating that the user's important locations were found 88% of the time, within 3 miles. Part of the evaluation was also made comparing it to other approaches used to estimate commute distance (Oracle, Top Two, and TimeBased). After this process, they showed how the algorithms can be applied to large-scale data analysis and policy planning by estimating the commuting carbon footprints of the population.

Yang et al. [39], developed an approach, based on criteria, to select significant places, without assuming that all users have only one home or one work. First, they proposed a gradient-based method to mine the user's trajectory by defining moving and stopping states, and then, they applied a method for noise handling. This phenomenon disturbs the trajectory mining results: sometimes, in the real world, the nearest antenna is not the one that serves the call due to load sharing. It affected 50% of the dataset, looking like the user moved kilometers in seconds. To deal with this noise, they applied the clustering algorithm Density Based Spatial Clustering of Applications with Noise (DBSCAN). Thus, they eliminated a large gradient of noise, caused by the re-selection, by clustering nearby antennas within a distance threshold *Epsilon (eps)* and set the cluster's center as the new location for all records in the cluster.

To determine significant places, they assumed that most of the people spend the night and dawn at home, so calls between 7PM and 5AM were made from home and daytime is spent at work, so from 8AM to 7PM most of the calls were made from the workplace. Hence, they applied these criteria on their dataset, containing registers from users in China, defining them as "Home Time" and "Work Time". Later, DBSCAN was also used to cluster all stops in a user's multi-day trajectory, identifying significant places.

Once they identified the places, they used it to calculate commuting distances and determine the density of home and workplaces distribution. The drawback in this work is that no ground-truth was used to validate the results and the evaluation was made by comparing their approach with another authors' approach which also did not validate their work.

Vanhoof et al. [36] used a CDRs dataset with registers of subscribers from France, to detect home locations and at the same time study the performance of five popular criteria used for home detection. Besides the comparison of the algorithms, this study also showed how spatial uncertainties at the individual level can be assessed in the absence of ground-truth annotations. In the first phase, they started by investigating the effects of five different criteria used in current Home Detection Algorithms (HDAs) to compare findings after validating them. The following criteria was later incorporated to construct five different algorithms. They define home as the place where:

1. The majority of both outgoing and incoming calls and texts is made;

2. The maximal number of distinct days with phone activities - both outgoing and incoming calls and texts - is observed;

3. Most phone activities are recorded from 7PM to 9AM;

4. Most phone activities are recorded, implementing a spatial perimeter of 1000 meter around a cell-tower that aggregates all activities;

5. The combination of 3) and 4) thus most phone activities recorded during 7PM and 9AM and implementing a spatial perimeter of 1000 meter.

The authors took several experiments into account, some of them using popular algorithms, like Hartigan's clustering algorithm and K-means, and constructed different algorithms, each incorporating a specific criterion with the aim of studying their performance and capabilities. They validated the results at cell-tower level, by comparing it with the distribution of homes per cell polygon based on data from the census. Proceeding, they constructed a framework to create a data-driven assessment of spatial uncertainty for home detection of individual traces and calculated the different measures for spatial uncertainty for all of the five HDAs, to investigate their temporal and spatial properties. To conclude they found the correlations between the spatial uncertainty measure and the validation results for all algorithms.

In some cases, these approaches were able of identifying, for each user, a possible home, second home, and even a third home location. After the evaluation and the comparison of the results, they concluded that the algorithms incorporating criteria of time constraints (3 and 5) disagreed with algorithms that count the number of activities (algorithm 1), distinct days (algorithm 2), or perform spatial grouping (algorithm 3), all of which showed accordance. The criteria of time constraints resulted in different detected places of residence for 30% to 40% of the cases compared to all other criteria. They assumed that these differences probably come from sparser observations and different spatial behavior during the night.

While publishing this article to evaluate the performance of the five HDAs, the same authors developed another study named "Assessing the Quality of Home Detection from Mobile Phone Data for Official Statistics" [35]. They used the results of the five HDAs to show that the efforts to detect home locations suffer from blind deployment of criteria to define the place of residence and, most of all, from limited validation possibilities.

In this work, it was argued that, over time, the approaches for home detection have been simplified to single-step, using decision rules based on simple *a priori* criteria to define a home location (as shown on the literature review above). With the both studies, they concluded that, when looking at the individual level, home detection methods are very sensitive to the criteria selected, and at high-level validation, the five algorithms performed with similar sensitive. Also, they accomplished that there is a large mismatch between population counts constructed from mobile phone data on home location and the validation dataset based on census data. The mismatch is unnoticed when only high-level validation is undertaken.

Although being common the application of rules to infer about home and work locations, their appliance implies that subscribers with different routines are treated the same way as the subscribers that have common routines. With the objective of identifying important places for different types of users, with distinct habits, some researchers did not make any *prior* assumptions on the behavior of the users.

An interesting contribution without criteria, is from Mamei et al. [25], that presented an approach to automatically identify places that people routinely frequent, using mobile network data produced by costumers of a telecom company from Italy. This work started by proposing the improvement of the state-of-the-art in the three phases that are typically adopted to determine important places, which are clustering, weighing, and thresholding.

Following their objective, they started by collecting CDRs of each subscriber and then clustered it in well specified spatial regions, taking into account the distance between the

registered CDRs. Algorithms that require as input the number of clusters to be found, such as K-Means, were found inadequate. Thus, in an early stage of the clustering process, DBSCAN was adopted, however, forward on the project was switched to a complex mechanism, an agglomerative algorithm, to constrain the size of the clusters. The next step was related to the weighing process, where they started weighing based on some factors (e.g., number of days in which the user visits the cluster) and then, they defined the threshold. Hence, clusters with a weight greater than a certain threshold were associated to relevant places.

To validate the research, ground-truth information, coming from a fraction of users, was used. This process consisted of collecting the number of places that the algorithm produced as result, for each volunteer user, and then, select the place (if any) that was closest to the ground-truth. If the distance between the place and the ground-truth was below a certain threshold (defined by the radius of the antennas coverage on the area), they considered to have found the place correctly. The evaluation process showed that the model found the locations of residence with 90% - 100% of precision and work locations with 80% of precision. Also, the correlation coefficient between the number of people living in a city, measured by the approach and by the Italian census was 0.91.

The paper of Burkhard et al. [6] is also remarkable. They proposed an approach to extract regular mobility patterns from sparse CDRs without *a priori* assumptions. This work presents two methods: one based on association mining and not computationally intense and another, named "DAMOCLES", based on extracting idiosyncratic daily patterns from clustered daily activities. As explained in section 2.1, these authors, used a technique to divide the day into equal time slots, where each one represented a CDR register. The first method was transferred from the GPS context and consisted of mining of association rules using an algorithm [15] combined with the cell ID and time slot as input. The "DAMOCLES" was the second method developed. It was based on clustering, and attempted to compensate low CDRs counts with aggregation over time, in a way that similar days were used to create prototype days. The algorithm integrated in this approach was divided into a dissimilarity measure, a clustering algorithm, and the reconstruction based on the identified clusters. Like other works, in this work DBSCAN was adopted to perform the clustering, as it allows a different number of identified clusters per user and accommodates the ones with different number of recorded days.

Both of the proposed methods were compared with two benchmarks from GPS measurements: one that assigns the most frequent cell, registered by time slot, to the time slots with no observations and another which assigns the mode of the cells observed by time slot and an indicator function for weekends. The results showed that the first method is faster and presents stable results, however, contrary to the "DEMOCLES", it is not able to capture the spatio-temporal information of the underlying data. Apart from this benefit, "DAMOCLES" presented limitations: due to the clustering, the method does not work for users with constantly very low numbers of events.

"Exploring Home and Work Locations in a City from Mobile Phone Data" is the title of a very interesting paper produced by Lumpsum et al. [34]. The approach that is following described, although not entirely, was a great reference to our project. The objective of the authors was the identification of the places of residence and work, through the determination of a sleeping period. To begin, due to the density of antennas in different areas, they represented the coverage areas of cell sites with a Voronoi diagram (urban areas have smaller polygons) and after they transformed the diagram into a regular grid with fixed size $500x500m^2$. In the next step, they developed two different algorithms: one for the identification of work location and the other for home location.

The method to identify workplace locations was focused on people whose workplace and working hours are fixed and consisted of finding a grid cell where the individual used the phone between working hours most regularly in terms of days and hour slots. To identify home locations, they started by identifying a sleeping period, based on finding a constant and long stop state across 24 hours. In other words, they excluded users with low activity and users whose daily usage distribution graph was not a U shape period. Then, they identified the hour slot with the minimum usage. If the client had multiple slots with minimum usage, the algorithm selects the the first one after 9PM. Once they found the hour that represents the sleeping hour, the events on the sleeping hour and +/- 4 hours were used to determine the home location.

The validation and evaluation of the results for home location were made only to the results of the home locations, by comparing the home grid cell with the sub-district address of the post-paid users. After the evaluation they realized that were able of identifying home locations for 34.57% users on the dataset with an accuracy of 69.02%.

2.3 Customer Segmentation

The above section (2.2) detailed some of the relevant works that were developed to identify home and work locations using CDRs. As noticed, the analysis of mobile data has grown into a mature research and became a research field with a wide array of applications. Although the mentioned works addressed the determination of meaningful places, none of them had their focus on doing that through the generation of profiles. Only the work developed by Lumpsum et al. [34] based their research for home and work locations on the characterization of users, according to their registers on the mobile phone network.

The generation of profiles, is a procedure that several companies adopt to characterize their clients. Usually, the scope of relationship marketing to retain clients is, above all by winning their loyalty. Once a person is connected to a company, the goal is to fulfill the person's demands and maintain the attachment. A common practice taken by companies is the process of dividing customers into discrete groups based on their common characteristics. This procedure is called customer or market segmentation and is seen by companies as an expansion stage, a way to scale efficiently, and effectively focus their efforts on specific subsets of customers [29]. This procedure triggered some authors' attention and a vital number of articles are dedicated to study the costumer segmentation in several areas [28].

The work developed by Băcilă et al. [26] used K-Means algorithm to segment prepaid telecom subscribers according the sum of amounts recharged, the value of SMSs sent, the Internet traffic value, and the value of calls over a period of six months. This segmentation algorithm (K-Means) follows the partition of the population based on their behavioral values and they are clustered in a way that the variation inside the groups is down to a minimum.

After completing the procedure, they identified seven segments inside the population. To determine the differences between the clusters, they used the ANOVA test, which indicated significant differences, among the identified groups, taken into consideration the case of each variable. The Tukey post-hoc test confirmed that averages of the four variables that were used as base of the segmentation, differ in the case of the seven clusters. The analysis indicated that only two groups of subscribers, used their credit for all types of services, and that 19.9% of the subscribers spent less than six monetary units on calls, 33.3% spent less than six monetary units on SMSs, and 23.6% spent below three six monetary units to connect to the internet.

2.4 Spatial Data Clustering

Clustering has multiple applications in various areas and comprises several approaches to group data, which define how clusters are created. This section gives an introduction of the clustering approaches found in the literature and some of the algorithms associated with each one of them. It is given special attention to the algorithms mentioned in the previous sections, such as DBSCAN, which was used in several works to identify meaningful places, K-Means that was used for segmentation processes in many papers [28], and to Varied Density Based Spatial Clustering of Applications with Noise (VDBSCAN) which was not mentioned above, however, it is a DBSCAN extension and an asset to this work. Then, the last part of this section is dedicated to a brief comparison between the three algorithms, concluding the state-of-art research.

It is observed in the literature that almost every model to identify land use, mobility patterns, or possible important locations use a data mining technique, called clustering, which can also be used to segment customers in several areas. In data mining, clustering is a discovery process that divides a population or a set of data points into several groups, in a way that promotes intra-group similarity and minimizes inter-group similarity [19].

Data clustering is a process that is usually taken to find patterns, points, or objects by partitioning or outlier detection, and it is also applicable to cluster spatial data. Objects within a valid cluster are more similar to each other than the objects outside the cluster [27].

Xi et al. [38] defined spatial data mining as the process of discovering patterns from large spatial databases. This procedure aims to automate the process of discovering patterns and involves spatial clustering, which is used to organize the set of spatial objects into groups (clusters).

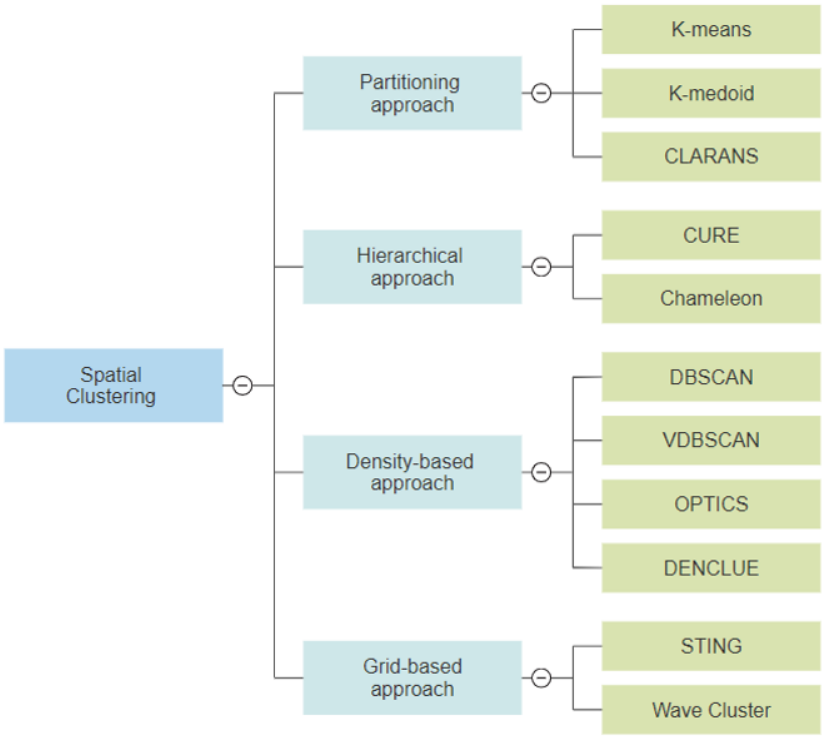


Figure 2.1: List of Spatial Clustering Algorithms (Adapted from Xi et al. [38])

Spatial clustering has been extensively studied and applied in numerous approaches developed by several researchers. Figure 2.1, was adapted from an article published by Xi et al. [38] ("Spatial Clustering Algorithms and Quality Assessment") and presents a classified list of spatial clustering algorithms into different approaches. Overall, these algorithms detect regions with higher and higher-than-expect rates, discover clusters with various shapes, and perform well with large datasets [11].

2.4.1 Partitioning Based

Clustering by partitioning is a process that runs iteratively and separates N objects into k groups. The first step is made by selecting k random points, and the objects are separated into k groups based on the distances between each point to the center. Then, uses the mean of each group or the nearest distance from the center to separate the objects into k groups [38].

K-Means

One of the most widely used and studied algorithms that belong to the partitioning approach is K-Means. The algorithm attempts to partition the dataset into k pre-defined distinct non-overlapping clusters where each data point belongs to only one group [10]. Given a dataset $D = \{p_i | i = 1...n\}$, p_i in d-dimensional space d, K-Means tries to assign the set of points into k clusters with arbitrary selected k initial centers. Depending on the final goal, K-Means divides a large cluster into several small groups or merges small adjacent clusters into a larger one, to achieve the minimal Sum of Squared Error (SSE). Formula 2.1 shows how the SSE is calculated, where $||p_i$ - $m_j||$ is the distance from point p_i to cluster center m_j, δ_{ij} is the cluster indicator variable with $\delta_{ij} = 1$ if $p_i \in C_j$ and 0 otherwise, and m_j is the mean of cluster C_j.

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n} \delta_{ij} ||p_i - m_j||^2 \tag{2.1}$$

Figure 2.2 illustrates how the algorithm clusters a set of data points with an pre-defined $k = 2$. Each colored point was associated to a cluster (purple and yellow), which are represented by their centroid (red circles).

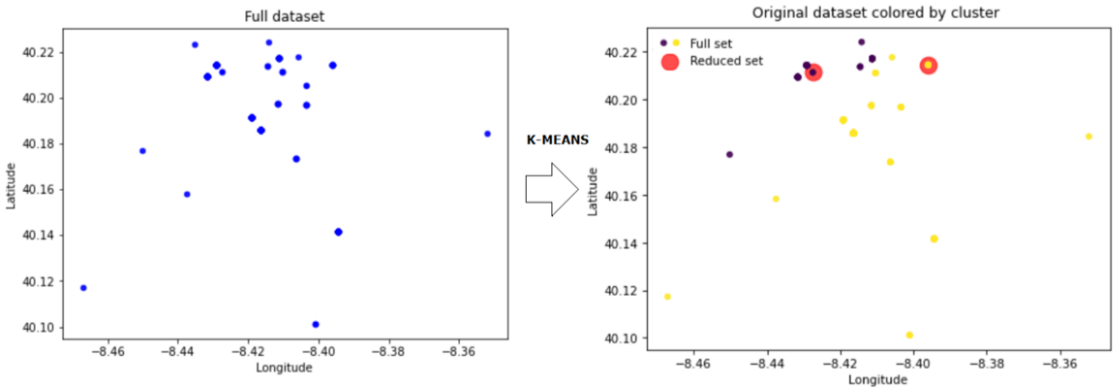


Figure 2.2: K-Means Clustering

Given to its simplicity, K-Means has been widely accepted. However, the algorithm suffers from a serious limitation: random initial centers selection may lead to getting trapped in poor local minimal [30]. Many research attempt to address the sensitivity of the initialization of this algorithm. One of these [3], proposed a careful seeds selection method, named K-Means++, however, in this technique, it is required to take some passes over the data to select the initial centers, which limits the application of the algorithm on large datasets. Besides this problem of sensitiveness, K-Means is also very sensitive to "dirty and abnormal" data and although there are other and most recent studies that attempt to solve this sensitivity problem [30], the improvement of quality and efficiency of the algorithm still is a widely studied area.

K-Medoids

K-Medoids is also a partitioned algorithm that attempts to minimize the dissimilarities between the points in a cluster and the point that is designated as the center of cluster. This algorithm is a classical partitioning technique of clustering that clusters a dataset, $D = \{p_i | i = 1...n\}$, of n objects into k clusters known *a priori*. The main differences between this algorithm and K-Means, is that K-Medoids chooses one of the original points as representative instead of choosing the cluster mean; it is a more robust algorithm to noise and not as sensitive as K-Means. Even though this algorithm can sometimes present better results than K-Means, due to its larger calculations, it is generally suitable for small datasets [7].

CLARANS

Clustering Large Applications based on RANdomized Search (CLARANS) is considered an improved K-Medoids and was initially developed for large datasets. However, its disadvantage is that it assumes that all objects to be clustered can reside in the main memory at the same time, which is not possible if the database is to large [27].

2.4.2 Hierarchical Based

Hierarchical clustering aims to build a cluster hierarchy where every cluster node contains child clusters (hierarchical tree). This approach allows exploring data on different levels of granularity and the clustering methods are categorized into agglomerative and divisive processes. The agglomerative clustering process starts with one-point clusters and recursively merges two or more most appropriated clusters. In contrary, the divisive clustering process starts with one cluster including all data points and recursively splits it into most appropriated clusters. Both these processes only stop when a stopping criterion, like a request of k clusters, is achieved [38]. It is very important that the criterion is well defined and not vague, or this can affect the algorithm's performance.

CURE

Clustering Using REpresentatives (CURE) is based on the hierarchical approach. In this algorithm, each cluster is represented by a fixed number of representative points. This number of points is reduced by a shrinking factor towards the center of the cluster enabling them to be free from outliers (noise). It is also able to handle massive amount

of data, which in a certain way, gives it an advantage over the algorithms based on the partitioning approach (Section 2.4.1). However, one of its disadvantages can be the "run time" complexity, $O(n^2 logn)$, which for certain Database Management Systems (DBMs) can be a high complexity to apply directly to large datasets [21].

CHAMELEON

CHAMELEON is a hierarchical clustering algorithm that operates on a sparse graph in which, the nodes represent data items and weighted edges represent similarities among the data items. In theory, these characteristics should allow it to operate with large datasets. CHAMELEON has two phases of action: first, it uses a graph partitioning algorithm to cluster the data items into a large number of relatively small sub-clusters; then, it uses an agglomerative hierarchical clustering algorithm to find the genuine clusters by repeatedly combining together these sub-clusters [19].

2.4.3 Density Based

Density-based approaches suggest that it is possible to recognize clusters because, the density of points is considerably higher inside the cluster compared to the outside. Thus, cells with high counts of points can be potential clustering centers [12]. One advantage of using density-based approaches is that with them, it is possible to find clusters of arbitrary shapes and sizes [38].

DBSCAN

DBSCAN is a density-based algorithm widely used for spatial clustering. In this algorithm, it is assumed that all points within a cluster are density reachable and points across different clusters are not [19]. The idea is that the neighborhood of a given radius must contain at least a minimum number of points. This means that, as in other density-based algorithms, two input parameters are required: the maximum physical distance between two samples, for one to be considered a neighbor of the other (*eps*) and the minimum cluster size, which is the number of samples in a neighborhood for a point to be considered as a core point, including the point itself (*MinPts*). In this way, the shape of a neighborhood is determined by the choice of a distance function for two points p and q, denoted by $dist(p, q)$. The neighborhood between two points p and q, can be explained by the following definition [12]:

$$N_{Eps}(p) = \{q \in D | dist(p, q) \leq Eps\} \tag{2.2}$$

In formula 2.2, p and q are considered neighbors if, in a set of points D, the distance between p and q is on the *eps* threshold. Thereby, the algorithm has the capability of clustering spatial datasets based on the radius, defined by the *eps* value and the minimum number of objects required, the *MinPts*. It also distinguishes three types of points that are illustrated in figure 2.3. This points are [40]:

- Core: A point is considered a core point if there are at least a *MinPts* number of points in the surrounding *eps* area;
- Border: A point is considered a border point if it is reachable from a core point;

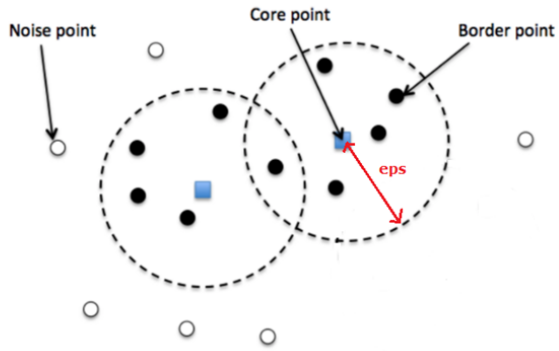


Figure 2.3: DBSCAN - Types of Points (Adapted from "DBSCAN Clustering Algorithm in Machine Learning")

- Outlier: A point is an outlier or a noise point if it is not reachable from any core points.

Figure 2.4, illustrates how the algorithm clusters data points with the an *eps*= 0.001 (degrees which corresponds to approximately 110m) and *MinPts*= 5. The figure shows that the algorithm found one group/cluster of data points based on the density of their distribution. The cluster is represented by the red circle, the data points that are in the area defined by the *eps* are allocated to the cluster (yellow), and the remaining ones are considered outliers (purple).

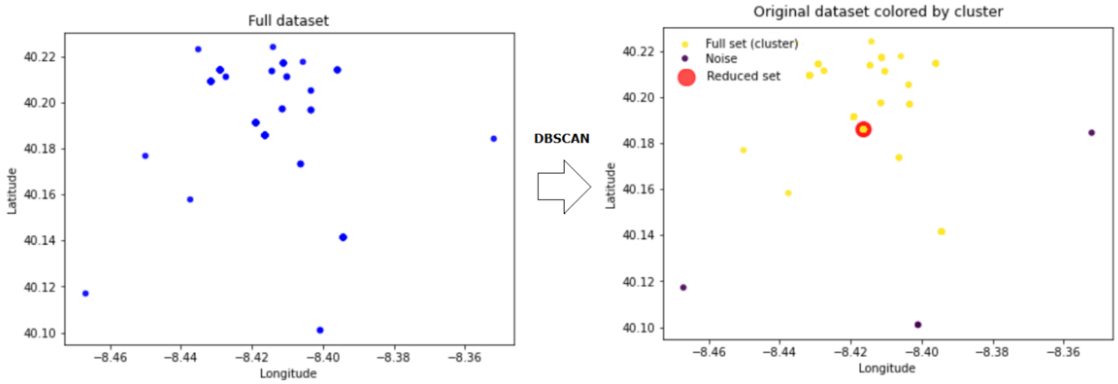


Figure 2.4: DBSCAN - Clustering

As with other density-based algorithms, with this one, clusters are easy to understand and their shapes or sizes are not limited. Yet, it may have trouble with clusters of varying densities [23].

VDBSCAN

In an ideal scenario for DBSCAN, the optimal input parameters are inserted and clusters are formed. As an attempt to accomplished that, VDBSCAN a variant of DBSCAN, emerged and proposes a method to analyze datasets with varied densities. The idea of this algorithm is the adoption of methods to select suitable values of the *eps* input parameter, before implementing the traditional DBSCAN. With different values of the parameters, it is possible to distinguish clusters with different densities simultaneously. The selection of

the optimal values is made by looking at the distance from a point to its k^{th} neighbor, which is called k-dist and results in a k-dist plot. The k-dists are computed for all data points for some k, and then the values of the distances are sorted in ascending order and plotted [23].

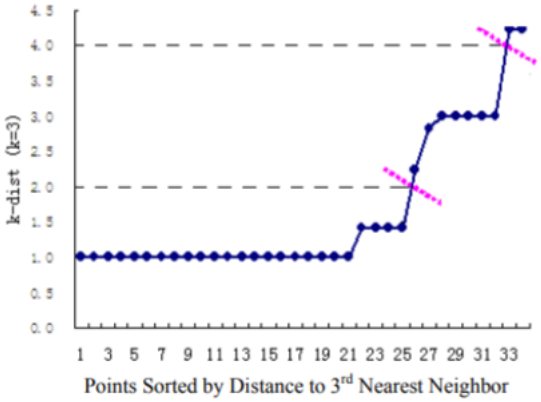


Figure 2.5: k-dist plot (Adapted from "VDBSCAN: Varied Density Based Spatial Clustering of Applications with Noise" [23])

In the k-dist plots, the sharp change at the k-dists value, corresponds to a suitable value of *eps*. Figure 2.5, was adapted from Peng et al. [23] article and presents the calculation of distances from each point to its 3^{rd} nearest point. After sorting the distances, the graph shows two suitable values of *eps*: $Eps_1 = 2$ and $Eps_2 = 4$. After this achievement the traditional DBSCAN is adopted twice for the two different values of the parameter. At the end of each iteration, marked-points will not be processed again. After all iterations, non-marked points are recognized as outliers [23].

Figure 2.6 shows how VDBSCAN clustered the same dataset that is presented in figure 2.4. Contrary to DBSCAN, we have only defined the *MinPts*= 5. The *eps* parameter was automatically defined by the k-dist plot in figure 2.7. This plot returned an optimal *eps*= 0.00546 (degrees which corresponds to approximately 607m), which was applied to the traditional DBSCAN algorithm.

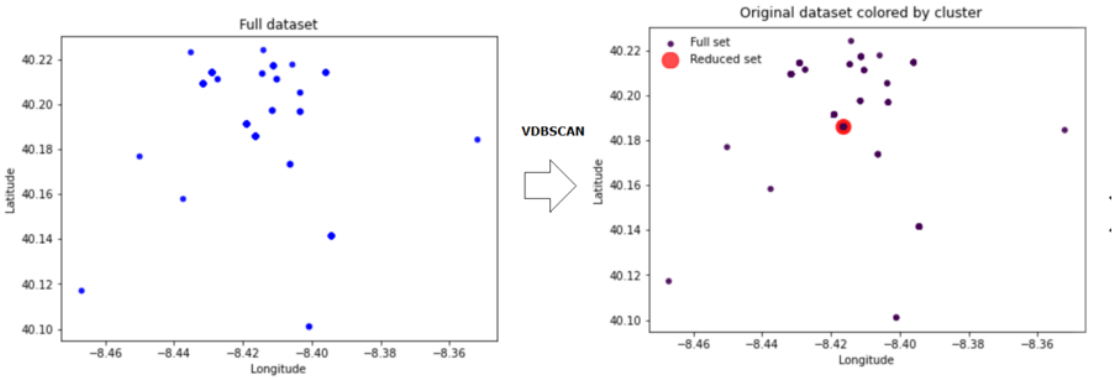


Figure 2.6: VDBSCAN - Clustering

Contrary to what occurred in DBSCAN, with the usage of an optimal *eps*, all data points (purple) were allocated to a cluster (represented by the red circle), showing that the choice of this parameter, without considering the sparsity of the points in the dataset, affects the quality of the results.

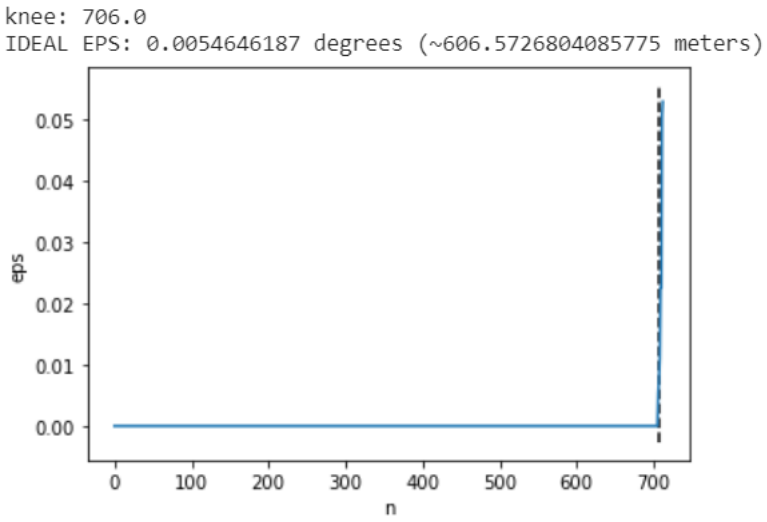


Figure 2.7: Optimal *eps*

OPTICS

Ordering Points To Identify the Clustering Structure (OPTICS) is an extension from DBSCAN, but simpler in terms of the required parameters. It creates an ordering of a database and stores the core-distance and a suitable reachability-distance for each object. A clustering structure is created, which defines a broad range of possible values and it automatically and interactively clusters the data [27].

DENCLUE

Density-based clustering (DENCLUE) aggregates partitioning and hierarchical clustering approaches. For this reason, it is considered more effective than other approaches. It works on arbitrary noise levels and on high-dimensional multimedia datasets, where other algorithms are not able to work. DENCLUE can be considered an improvement of DBSCAN and there are cases where performs much better than this algorithm [27]. However, such as DBSCAN can have trouble with density if the density of clusters widely varies [38].

2.4.4 Grid Based

The grid-based approach can contain both partitioning and hierarchical algorithms but differs from them because the algorithms are not focused on the data points, but on the value space that surrounds them.

STING

Statistical Information Grid Approach (STING) is considered a grid-based algorithm because it retrieves spatial data and decomposes it into several cells using a rectangular hierarchical structure. Then the mean, variance, minimum, and maximum of each cell are computed and a grid structure is formed, where the new objects are inserted. It is one of the highly scalable algorithms that can decompose the dataset into several levels of detail [27]. It also assembles statistics in a hierarchical tree of nodes that are grid-cells.

Wave Cluster

Like STING, Wave Cluster decomposes the data into different levels of hierarchy. It uses a technique, based on signal processing, that decomposes a signal into different frequency sub-band. The data is transformed to preserve the relative distance between objects at different levels of resolution. This process allows natural clusters to become more distinguishable. It is highly scalable and can handle outliers well, but is not suitable for large datasets [27] and the process seems to be complex for daily research in this area.

2.5 Conclusion

After describing the various algorithms that belong to different approaches, with specific attention on the ones that were used by other authors to identify important places and segment clients, we are going to make a brief comparison between K-Means, DBSCAN, and VDBSCAN, which are the ones relevant for our work.

As described and observed in the sections above, the K-means and DBSCAN and VDBSCAN cluster differently. While K-Means clusters the events by similarity and it is necessary the definition of the number of clusters to be formed (k), DBSCAN and VDBSCAN form clusters based on the density of events (defined by the input parameters). Table 2.3 presents the behavior of the algorithms regarding to the clusters shape, the necessary input parameters, and how they deal with outliers and with different densities. This comparison between them is used to highlight their differences.

Table 2.3: K-Means *vs* DBSCAN *vs* VDBSCAN

	K-Means	DBSCAN	VDBSCAN
Clusters shape	Clusters are more or less spherical or convex in shape and must have the same feature size	Forms clusters with arbitrary shapes	Forms clusters with arbitrary shapes
Input parameters	The number of clusters is an input parameter	The number of clusters does not need to be specified, but the global parameters (*eps* and *MinPts*) need to be specified	The number of clusters does not need to be specified and only the *MinPts* parameter need to be specified
Outliers	Very sensitive to outliers and noisy datasets	Efficient at handling outliers	Efficient at handling outliers
Densities	Varying densities of the data points does not affect the clustering	Trouble when datasets are sparse or data points have varying density	Sparse data or data points with varying density can be handled by the automatic generation of several *eps* parameters

After studying the algorithms and comparing them, it is clear why some authors adopted density-based algorithms to identify important places. Although K-Means is commonly used for clustering, when regarding spatial data, and attending to our purposes, DBSCAN can be superior. Besides discarding sporadic events, by considering them outliers, the algorithm is based on the density of events, separating areas of high density from areas

of low density. Furthermore, DBSCAN figures out the number of clusters automatically, which allows, for example, the identification of second home locations. Moreover, the algorithm handles clusters of various shapes and sizes, as opposed other algorithms, like K-Means, which works best for convex shapes.

However, the usage of VDBSCAN, will also guarantee the attribution of an optimized *eps* for clients who live in urban areas and clients who live in rural areas (urban areas have a higher density of antennas) allowing the correct identification of the important places, independently of being in a urban or rural environment.

Regarding the characterization of users, K-Means seems to be the ideal algorithm to perform their segmentation into adequate groups. If we use the right methods to optimize the number of clusters (k), and the adequate variables to be the base of the segmentation, the clustering process can be efficient. Later, after this segmentation, the features of each group can be deeply analyzed and the clients can be profiled.

In the literature described in the previous sections, it is noticed that some researchers used *a priori* assumptions to identify significant places. These assumptions were transferred to the models through the application of criteria (most of them temporal criteria) and applied to entire datasets without exception. This means that, for example, when the authors assumed that home locations are the place frequented during the night [16] [39] [36], for users that work during the night, their home locations were identified at their work location and vice-versa. By segmenting clients and analyzing their profiles, the right criteria to identify home and work locations may be applied to each group.

Chapter 3

Data Analysis and Processing

Before starting the implementation, it is important that the data is cleaned and pre-processed, as the absence of these steps can result in a profound impact on the performance and results of the model.

This chapter introduces and describes the datasets collected and provided by a telecom company. It also addresses the steps taken to analyze and treat the data. The tools necessary to perform this analysis are described in the next chapter (section 4.2).

3.1 Datasets: Antennas, CDRs, Snapshots, and Ground-truth

Along with the project development, the telecom company has provided us registers of CDRs and Snapshots of their subscribers from June to October of 2020. These registers belong to 35 831 clients with most of their registers recorded in the antennas of Coimbra's district. Both of the data types are described in Section 2.1, where is also presented some of the columns that the data may contain. However, as explained in that same section, operators can include, remove, and tag fields as they choose.

The data was provided in separated Comma-Separated Values (csv) files: a file containing the information referent to the location of all cell towers where the antennas of the company are allocated in Portugal and other files with the records of the subscribers. The goal of this analysis is the elimination of columns that are irrelevant to this project and the creation of a unique and organized file containing all the data resulting from the filtering.

3.1.1 Antennas

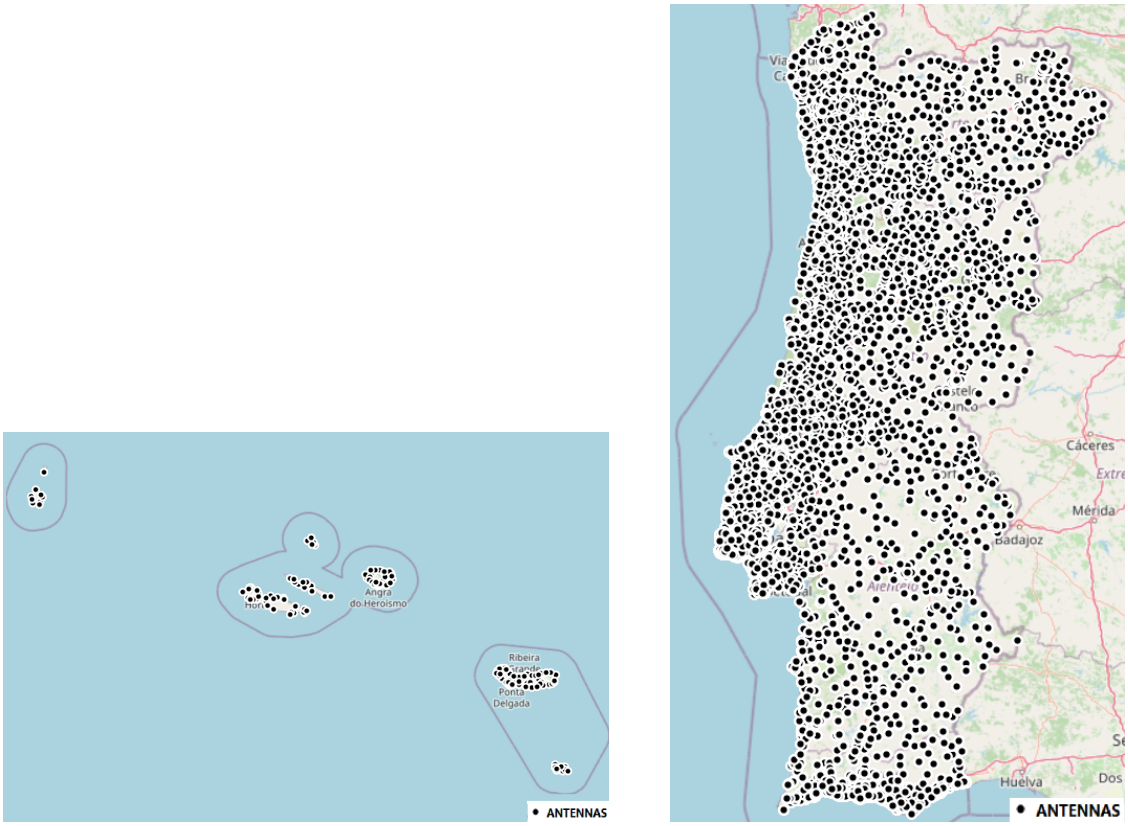
An important procedure is the analysis of the distribution of antennas among the country and, especially, the concrete area that is being studied. Table 3.1 presents a sample of registers of the dataset accommodating information relative to the antennas.

Table 3.1: Antennas

cellID	LAT	LONG	PARISH	COUNTY	DISTRICT	COV(m)
7566	41.538472	-8.525278	Martim	BARCELOS	BRAGA	1191
2667	40.652372	-7.889041	Ranhados	VISEU	VISEU	4900
22766	39.675626	-8.532025	Seiça	OURÉM	SANTARÉM	2400

In table 3.1, the "cellID" field contains a unique number that identifies each antenna, the "LAT" and "LONG" fields give the exact geographic location, and "PARISH", "COUNTY", and "DISTRICT" describe the geographic location of the antenna at different levels. Finally, the "COV" field gives the coverage radius of the antenna (in meters).

Figure 3.1 presents the geographic distribution of the antennas over the entire country of Portugal, where it is notable the higher density of antennas in regions that are densely populated.



(a) Portugal Islands (b) Portugal Mainland

Figure 3.1: Antennas of Portugal

Table 3.2 presents the analysis of the number of antennas and their coverage radius. With this information, it is observable that some antennas are able of covering an entire village or even a district.

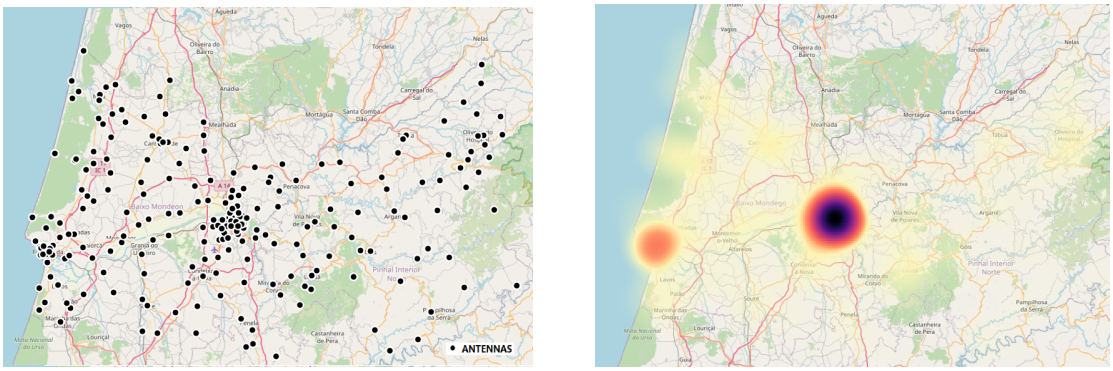
Table 3.2: Analysis of the antennas' radius of Portugal

Number of antennas	52 659
Mean of the radius	2 840m
Standard Deviation	2 996m
Minimum radius	60m
Maximum radius	32 000m

Antennas of the area studied (Coimbra)

Figure 3.2 is focused on the area that is being studied, presenting the antennas that are allocated in the district of Coimbra. In the illustration of the heat map in figure 3.2(b) it is

clearly visible the different densities of cell towers/antennas across the region, highlighting a higher density in the urban areas.



(a) District of Coimbra

(b) Heat map with the distribution of antennas

Figure 3.2: Antennas in Coimbra's district

The difference in the number of cellular infrastructures to support antennas between rural and urban areas is not only observable in Coimbra, but all over the country. Thus, this shows that regions with a higher dense population, have a higher quantity of antennas to support the users' needs, meaning that this distribution is like a photograph of the distribution of the community density across the country.

Following the strategy of Lumpsum et al. [34], we use a Voronoi diagram to visualize the estimated coverage area of each antenna. Figure 3.3 is an illustration of the coverage area of each antenna in theory (polygons) and just like in their work urban areas have smaller polygons.

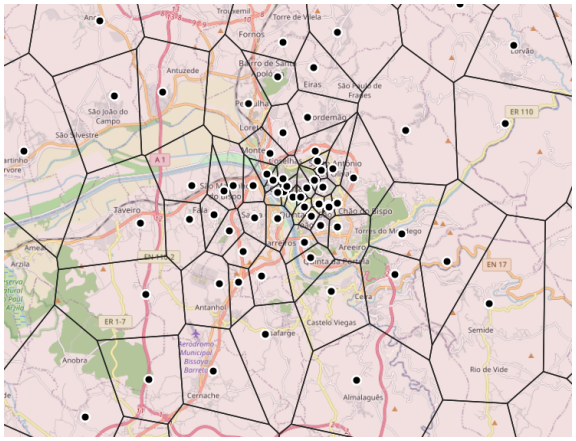


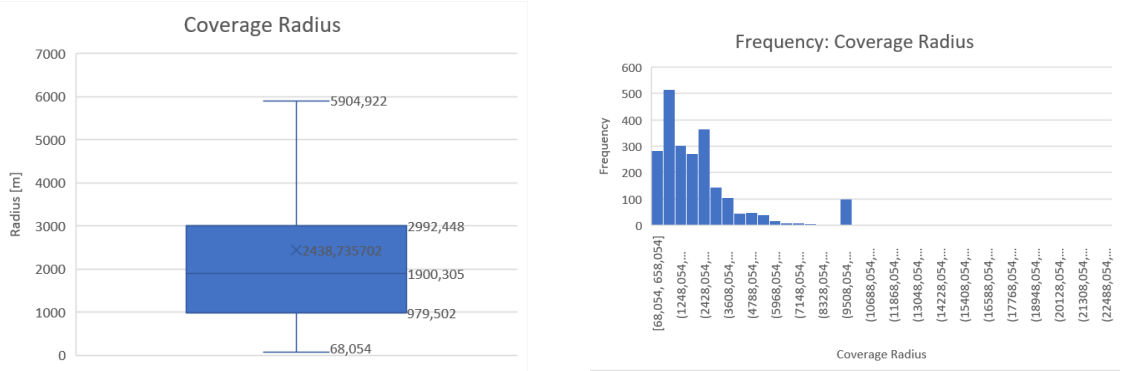
Figure 3.3: Antennas and their hypothetical cell coverage areas

Table 3.3 presents the analysis of the coverage radius of the antennas in Coimbra's district. It is observable that the antennas in Coimbra also have very distinct coverage areas between them. Figure 3.4 shows more clearly the analysis made, highlighting that besides the presence of antennas with an elevated coverage radius, most of them are below 5 905m.

So, as presented in tables 3.2 and 3.3, each antenna has its own coverage radius, and, in real life, the area covered by them is irregular and most of the time overlaps each other, which makes the diagram in figure 3.3 not credible. As analyzed, there are also some more complex cases where some coverage areas wrap another entirely.

Table 3.3: Analysis of the antennas' radius of Coimbra

Number of antennas	2 342
Mean of the radius	2 438m
Standard Deviation	2 227m
Minimum radius	68m
Maximum radius	23 000m



(a) Mean of the coverage radius registers

(b) Frequency of the radius coverage

Figure 3.4: Area coverage of the antennas in Coimbra

Due to the antennas' distribution, it is expected that the mean radius in urban areas is smaller compared with the mean coverage radius in rural areas.

3.1.2 Call Detail Records (CDRs)

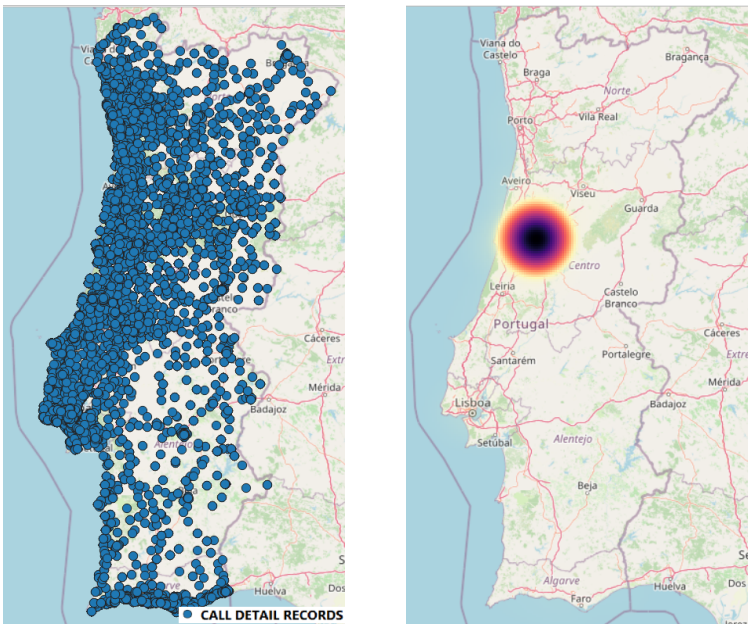
The datasets containing the CDRs throughout June to October comprehend anonimized registers of calls and text messages made by the subscribers. Table 3.4 exhibits a sample of the data.

Table 3.4: Sample of CDRs

DayCODE	HourCODE	MinSecCODE	UserID	CardCODE	CellID
20200901	0	5246	F9E21...	4829661	66
20200901	1	1636	3C52B...	475A46A	68
20201025	0	3608	D2FA0...	95107C4	32876

In table 3.4, the first three columns present the day, hour, and minutes and seconds, respectively, where the event took place. The "UserID" and "CardCODE" fields contain the anonymized identifiers that associate the register to the client. Lastly, the "CellID" field is the ID of the antenna where the event was registered, which allows the connection of this dataset with the antenna's dataset.

Figure 3.5 presents a sample of the registers on the CDRs dataset. Although the registers are from subscribers with more events in Coimbra, the illustration shows that they were recorded in antennas all around the country. Still, it is also visible in the heat map based on density, that the records have a higher incidence in the region of Coimbra.



(a) CDRs entries

(b) Heat map with the incidence region of the events

Figure 3.5: Sample of CDRs

3.1.3 Snapshots

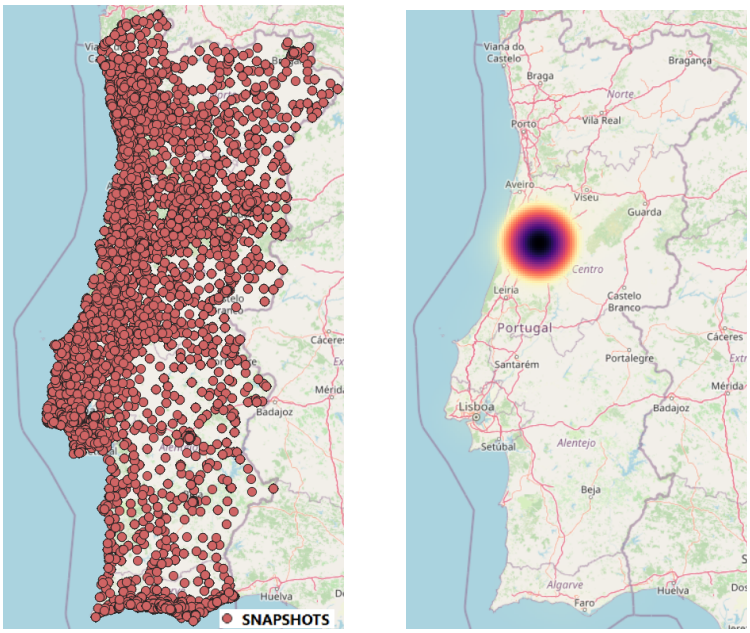
As explained in section 2.1, snapshots are a type of data that tries to avoid burstiness in the records. This means that it tries to mitigate the sparsity and irregularity that is, most of the time, observed in CDRs. Thus, instead of recording all events performed, only the last event made on every two-hour slot is recorded.

Table 3.5 presents a sample of the dataset provided by the company. This dataset is constituted by "Dtime" which is the timestamp, the "CellID" field (to assure the connection between this data and the antennas dataset), the "DUMPTS" field containing the identification of when the snapshot was documented, the "MSISDN" with the anonymized identification of the user, and the "UniqueCODE" with a unique code for all events registered, also anonymized.

Table 3.5: Sample of Snapshots

Dtime	CellID	DUMPTS	MSISDN	UniqueCODE
2020-09-03 21:32:20	573453	202009040800	100414...	A32798F...
2020-09-04 08:06:30	573451	202009040800	21C42...	34E49C4...
2020-09-18 19:35:58	1604374	202009182000	735FD...	25E126E...

Figure 3.6, illustrates how the snapshot events are distributed on the country's map. Similar to the CDRs, as expected, the events were registered in antennas all over the country, with a higher incidence in Coimbra's region.



(a) Snapshots entries

(b) Heat map with the incidence region of the events

Figure 3.6: Sample of Snapshots

3.1.4 Ground-truth

To this project only ground-truth to validate and evaluate the home location was available. The data to perform the validation was also provided by the telecom company associated to this project and the file contains information of the centroid of the postal-code of 4 600 post-paid subscribers. These subscribers are a set of users extracted from the previously mentioned files (CDRs and Snapshots datasets).

Table 3.6 presents a sample of the content of the ground-truth. The "UserID" and "UniqueCODE" columns contain the same information as in the CDRs and Snapshots files and allow the connection between the datasets.

Table 3.6: Sample of the ground-truth dataset

UserID	UniqueCODE	POTAL-CODE	LAT	LONG
0003F...	5C6029...	9125-239	32.654385	-16.831698
000E4...	0F708E...	3030-871	40.172332	-8.3918335
012B9...	B7F6B...	3040-382	40.199345	-8.4346753

3.2 Data Processing

The previous analysis allows the elimination of irrelevant columns. Since some of the columns, presenting the same content, are in different formats, for example, the timestamps, they must be treated in a way that in the end, are consistent between them. The final process of this phase is merging the resultant contents of the antennas, CDRs, and Snapshots files into a final dataset, which, when again processed, is going to be the basis of this project.

Table 3.7 contains a sample of the resultant dataset. First, the CDRs and Snapshots datasets were filtered, eliminating the unnecessary fields. Then, identical and necessary columns were aligned, and finally, the datasets were concatenated into a single file. The column with the month was separated from the date, for a matter of practice in future work, since different months are part of different seasons and are adequate to identify different places. A new field, "TYPE", was also added and is responsible for distinguishing snapshots from CDRs and events documented on workdays from events documented on weekends. This characterization is necessary to, for example, distinguish primary homes from second homes.

Table 3.7: Sample of the final dataset

UserID	CellID	LAT	LONG	Date	Hour	Month	TYPE
6D38B...	30439	40.250961	-8.43174799	2020-07-21	14	7	CDR-WORKDAY
EC7A7...	1946409	40.204103	-8.4189929	2020-10-05	22	10	SNAPSHOT-WORKDAY
4BD80...	609804	40.2234780	-8.43527714	2020-10-11	1	10	SNAPSHOT-WEEKEND

3.2.1 Resulting Dataset

The resultant file has a total of 42 612 929 records belonging to the 35 831 subscribers. One of the goals of this project is the attribution of geo-profiles which is only possible if there is enough information on the client. So, clients that do not have enough events to be profiled must be eliminated from the user base. Following Ahas et al. [2] work, it is assumed that clients with events in their most visited cell on fewer than seven days a month, are not suitable and their geo-profile is an impossible task to perform. So, they were excluded from the final dataset.

After restricting the user base, 41 371 218 records, 97.09% of the initial sample, belong to the resultant dataset, with a total of 5 102 users eliminated from the base and 30 729 remaining. After this process, the final dataset and the ground-truth dataset have 4 292 subscribers in common. It is also observable that the availability of the two types of data allows that more clients are characterized in geographical terms, because, users who did not have enough events of one data type to remain in the user base, with the join of the other type, can now be geo-profiled.

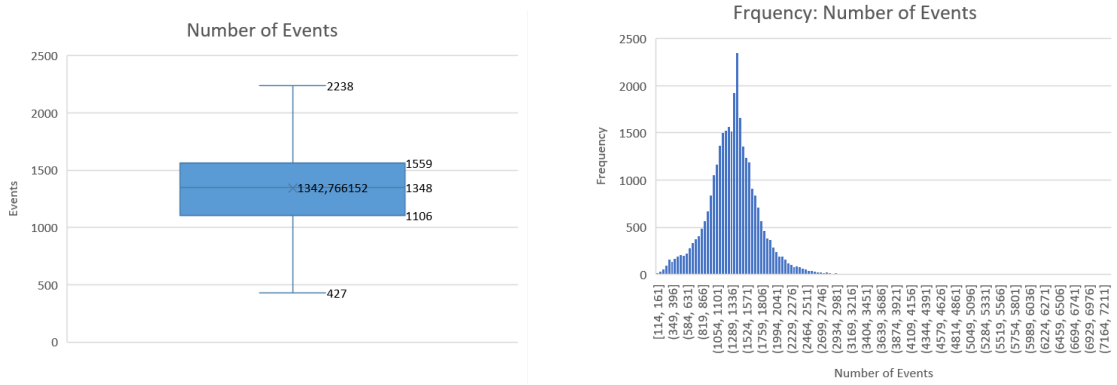
The quantity of data in the final dataset is representative and allows us to perform experiments and apply techniques to infer significant places and obtain accurate results. Also, since we are dealing with a considerable amount of records, we can study the algorithm's run-time complexity.

The first process of filtering is now concluded, however, the dataset still has records that are making reference to unknown antennas (table 3.8), which will be ignored when each client's dataset is individually treated.

Table 3.8: Sample of the undefined registers

UserID	CellID	LAT	LONG	Date	Hour	Month	TYPE
35CE2...	-1	-1.0	-1.0	2020-08-21	14	7	CDR-WEEKDAY

The analysis of the data is a crucial step to apply clustering algorithms in terms of complexity and understand their results and scalability. Since the geo-profiling is going to be made user by user, it is important to perform an analysis on the number of events of each user on the dataset.



(a) Mean number of events

(b) Frequency of the number events

Figure 3.7: Frequency on the number of events per user

Figure 3.7(a) presents the empirical distribution of the events in the dataset for each user, highlighting the average number of events per user. The outliers were excluded from the illustration, however, in the graph illustrated in 3.7(b) it is possible to observe that for the 30 729 users on the dataset, the number of events per user varies between 114 and 7 211 events.

Chapter 4

Geo-Profiling and Meaningful Places

Having analyzed, interpreted, and pre-processed the final dataset, it is time to implement the methodology to achieve the objectives. This chapter gives an overview of the methodology implemented to identify the users' meaningful places and geo-profile them. Throughout the chapter, two sections are describing how the project was conducted: the methods and techniques applied and the tools that were necessary to apply them.

4.1 Methodology

The necessary steps to implement the methodology are summarized in figure 4.1. It describes an iterative process, divided into several phases that were meticulously thought and designed to fulfill the expected outcomes. With this, the first phase, described in Chapter 2, was dedicated to studying the existing approaches and understanding the models, algorithms, and challenges.

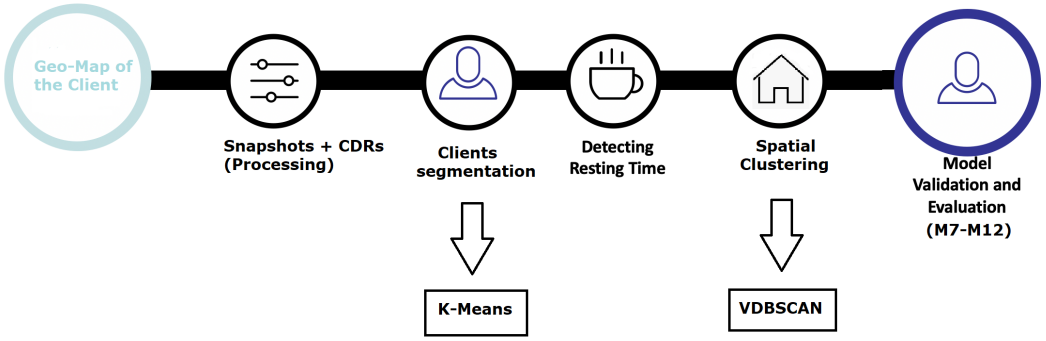


Figure 4.1: Methodology

As previously described in Chapter 3, the data was analyzed and prepared, sometimes by eliminating unnecessary columns or generating new necessary attributes. This phase is related to the pre-processing of the dataset, and from that resulted a dataset with only users that have the necessary events to be geo-profiled.

The next step consists of using a clustering algorithm, K-means, to segment the set of

costumers present on the final dataset, characterize the resultant groups, and infer on a period that is assumed to be part of their sleeping period.

Then, a density-based algorithm, VDBSCAN, is used to infer home and work locations of each user, at cell tower/antenna level, based on his/her routine features. This algorithm is also used to infer second home locations.

After collecting the outcomes, they are addressed to a spatial database, PostgreSQL, allowing their visualization in QGIS, a geographic information system. The final step consists of the validation and evaluation of the obtained results. If the validation or evaluation does not correspond to what is expected, then the model must be improved. Once the results are validated and evaluated, the model is applied to larger samples of clients to understand and study its scalability.

4.1.1 Costumer Segmentation and Sleeping Periods

After analyzing different techniques to determine sleeping periods, such as the determination of a sleeping hour (an hour with minimum activity) to determine the sleeping period, which was used by Lumpsum et al. [34] or the identification of a five-hour slot with less activity, we decided to adopt a technique that allows us to identify the users' routine features (their profile), the customer segmentation.

As discussed in the previous chapters, companies are competing to provide the quality of service demanded by the market. So they have created a stratagem to segment clients, in which they study the client's individual patterns, habits, demands, etc., and group them according to their individual characteristics. Then, clients are approached differently according to their needs/profile.

Identifying a period that is assumed to be part of the user's sleeping period is our first step to build the profile. Inspired by the companies stratagem, in this project, the 30 729 users present on the final dataset are segmented and grouped according to their interaction with the mobile phone network. This is a technique that was also used by Băcilă et al. [26] and consists of using a partitioning based algorithm, K-Means, that clusters by similarity, to segment the subscribers according to the number of events they have in different slots of the day throughout several days.

As mentioned in Chapter 2, K-Means is a clustering algorithm that attempts to partition N observations (the dataset) into k pre-defined clusters where each data point belongs to only one group. Before performing the segmentation, the base of the division has to be properly selected. In this case, the base is the behavior of the clients in terms of days and hour slots. However, the days of the week and the time slots must be accurately selected.

With K-Means it is necessary that the number of clusters is given before the clustering (it is an input parameter), and must be accurately selected. So, after defining the variables to perform the clustering, a method to optimize the input parameter is used. We will use the elbow or knee method, which is also used to gather the optimal *eps* in k-dist graphs (VDBSCAN). The elbow/knee was explained in Section 2.4.3 and, in K-Means case, consists of plotting the Within Cluster Sum Of Squares (WCSS) against the number of clusters to obtain the optimal quantity of clusters. For each k, the WCSS is calculated, determining the sum of squared distances between the observations and the cluster's centroid, which is

given by the formula:

$$WCSS = \sum_{i \in n} (X_i - Y_i)^2 \tag{4.1}$$

In formula 4.1, Y_i represents the centroid from the observation X_i. As the k values increases, the distance between the points and the centroid decreases, so, the plot looks like an elbow/knee. Like in the k-dist plots used to find the optimal *eps* parameter, the optimal k is the value where the curve abruptly changes and starts going almost parallel the X-axis [4]. Once the optimal input for the number of clusters is obtained, K-means is applied and the clients are grouped according to the similarity that exists between them, on the number of events in slots mentioned above.

In their study, Lumpsum et al. [34] identified an hour with less activity (sleeping hour) and declared the sleeping period as the period of four hours before and after the sleeping hour, resulting in a nine hour period. Yet, this period is a bit longer, to apply in this project, if the different groups have identical routines. Thus, after clustered, each group is analyzed, and according to its behavior, a five hour slot with less activity (sleeping hour and four hours adjacent), is inferred as the sleeping period.

The method to identify important places is focused on subscribers whose sleeping period and work hours are fixed. So, based on the identified sleeping period, the groups will be classified as night or day workers and according to this classification, also a five hour slot will be determined as the work hours. If there is a group where these slots are impossible to identify, the sleeping period and the work hours are going to be determined based on the typical hours to execute these activities (sleeping period during the night and work hours during the day).

Then, given the group that they belong to, the clients are labeled and a new column is added to the dataset, identifying the profile of the user associated with each event. Depending on the events made throughout the sleeping period and the work hours, the home and work locations, respectively, are identified. Basically, this process consists of attributing time constraints criteria, to each group, established by the analysis of the users' activities throughout the day, to identify the mentioned locations.

4.1.2 Spatial Clustering: Home, Second Home, and Work

Once the sleeping period of the user is inferred, it is possible to infer about home, second home, and work locations. So, for each user, a clustering algorithm is applied to group their events and identify important places. This identification is made using VDBSCAN, an extension of DBSCAN. VDBSCAN is a density based algorithm that identifies places where the user has a significant number of events and is able of dealing with datasets with varying densities. The selection of the algorithm was based on some articles presented in Chapter 2 that used DBSCAN to identify home and work locations [39] [6] (DBSCAN and VDBSCAN are described in Section 2.4.3).

The algorithm is applied three times, upon three different matrices to identify the three important places. The matrices contain the latitude and longitude values of the events, of the client that is being treated. One matrix contains events on the sleeping period to determine the home location, other contains events on the work hours to determine work location, and another contains events that allow the identification of second home locations. So, the home and work matrices are originated according with the label that

was assigned to the user on the previous process.

$$coords\_x = df\_x.[[LAT, LONG]].values \tag{4.2}$$

In formula 4.2, the x, in the $coords\_x$ and df_x, stands for "home", "sec_home", and "work" and the formula represents how the three matrices are originated ($coords\_home$, $coords\_sec\_home$, and $coords\_work$). The df_x corresponds to the dataframes with the events that, allow the home, second home, and work identification.

In an early stage of the project, before implementing VDBSCAN, DBSCAN was used with the same set of parameters to identify the different places, for all users in the dataset.

$$dbscan = DBSCAN(eps = 0.01, minsamples = 2).fit(coords\_x) \tag{4.3}$$

Formula 4.3 was used to implement the algorithm in Python with Sklearn library. The $eps = 0.01$ (degrees which corresponds to approximately 1.11km) was applied, because with the use of this input value (two decimal places for latitude and longitude) the algorithm is capable of recognizing the village where a user lives and works, and villages where possible second homes are located. The $MinPts = 2$ was applied because, according to Yang et al. [39], may re-select among two or more antennas in the area nearby. The metric used to identify the clusters was the Euclidean, which is the default metric to this library and measures the straight line distance between *coords* points (distances between nearest points/events). Due to the distances between events in consecutive hours not being significant most of the time, we use this metric instead of Haversine, which determines the great-circle distances between points taking into account the curvature of the earth.

As mentioned, this algorithm was used at the beginning of this project, however, after several experiments, the different densities of antennas in rural and urban areas made the use of this algorithm, a difficult task to perform, due to the fixed *eps* parameter. Also, with the different densities of events in each user's dataset, the parameters applied were not ideal, returning, sometimes, several locations for one important place and other times none. Nevertheless, this problem can be solve with VDBSCAN. As explained in Section 2.4.3, with this algorithm a technique to determine the optimal *eps* value is performed and then the traditional DBSCAN algorithm is applied.

So, user by user, each matrix is analysed, a k-dist plot is created and the optimal *eps* is returned. Next, this value is applied on the DBSCAN formula. Every time that is not possible to determine an ideal value of the parameter, sometimes because the events are all on the same location and the k-dist has no curvature, the *eps* is defined as 0.01 degrees. In a this phase, the Euclidean metric was maintained and the *MinPoints* parameters was also refined.

After selecting the right parameters the algorithm clusters the events and the antenna that is closest to the centroid of the cluster is identified as the important place.

Home

The purpose of determining a sleeping period, is to identify when the user's activity is minimal, because it is assumed that he/she is sleeping. The place where this minimum of activity takes place will be considered the home of user. When identified the sleeping period, the events registered throughout this period on the workdays from September to

October, are filtered and clustered. It is only needed the occurrence of one event during five hour slot and the event location is determined as home location.

$$dbscan = DBSCAN(eps = optimal, minsamples = 1).fit(coords\_home) \tag{4.4}$$

If there is no events throughout this period, the events on the two hours before and after, are collected and the same formula (4.4) is applied. In cases that even after this, there is no event registered, it is declared impossible to identify home locations.

Workplace

The work location is also identified based on the sleeping period. Given the label that was attributed, the work hours are inferred. This inference is made taking into account the Portuguese work system. After reading the user's label, the events made on workdays from September to October, on the work hours, are filtered and used to determine the workplace. To determine the work locations the following parameters were applied:

$$dbscan = DBSCAN(eps = optimal, minsamples = 5).fit(coords\_work) \tag{4.5}$$

The *MinPoints* = 5 was applied, because, besides being the default value of the library, with the five hour slots of work hours, it was assumed that the user must have at least five events on the five hours.

Second Home

The identification of second home locations does not rely on the sleeping period analysis and is done equally for all users. According to the Portuguese work system, the typical vacation periods are 15 or more days from the 1^{st} of May to the 31^{st} of October. So, to determine this place, it is assumed that second home locations are where the user has more than five events from 7PM to 7AM [16] on workdays and weekends from July to August, which in Portugal is the common vacation season, and on weekends in September and October.

$$dbscan = DBSCAN(eps = optimal, minsamples = 5).fit(coords\_sec\_home) \tag{4.6}$$

In this case, also the *MinPoints* = 5 was applied because it was the value applied to the workplace identification and for being the default value of the library.

4.1.3 Validation and Evaluation

After identifying the important places, it is important that they are validated and evaluated. Due to the lack of ground-truth available, in this project it was only possible to perform the validation and evaluation on the home location outcomes.

The validation and evaluation are based on the methods that Mamei et al. [25] followed to validate and evaluate their results. These authors, also collected information from a fraction of their users to validate the results, and to perform the evaluation compared the

distances between the place found and the ground-thruth. If the distance between the two places was below a certain threshold, they considered having found the place correctly. This threshold was adjusted according to the average radius of the cells covered by the antennas in the area (city or suburbs).

Inspired by the mentioned work, our evaluation method consists of using the coverage radius of the antennas to declare if the place was accurately found or not. As mentioned in Section 3.1.1 antennas have different coverage areas, besides that, the density of antennas placed in urban areas is a lot higher than in rural areas. Considering these differences, to evaluate the outcomes, areas with different densities of antennas are differentiated, to achieve the mean coverage radius in each one. To perform this distinction between the zones of the district of Coimbra, DBSCAN is once again applied.

Then, the mean radius of antennas in high dense (urban) and less dense (rural) areas is used to perform the evaluation. After inferring about the home location of the users that belong to ground-truth dataset, we use QGIS to measure the ellipsoidal distance, in a straight line, between the centroid of the postal-code, on the ground-truth dataset, and the location of the antenna that was identified as the home location. The ellipsoidal is the distance along a great circle on an ellipsoidal body.

The results are then evaluated in a way that if in a high or low density area, the distance between the two places is less than the coverage radius of the respective area, the home location is considered achieved with success.

These technique to validate and evaluate the accuracy of the results was adopted based on the available information that we have. Lumpsum et al. [34], evaluated their results on the home location by creating home grid-cells (based on a Voronoi diagram) and comparing them with the sub-district address of the clients. However, with the analysis performed on the antennas of the region that is being studied, we conclude that antennas with large coverage, cover multiple grid cells, which could cause noise in the results. Other authors, such as Isaacman et al. [16] and Vanhoof et al. [36] [35] used census to perform the validation and evaluation, however, the data on these surveys can be very outdated.

Also, Lumpsum et al. [34], only validated and evaluated the results of the home locations, because usually, the telecom companies do not have profile information about users' workplaces. Yet, is common that they have information on the address of post-paid clients.

4.2 Technologies and Tools

Various technologies were employed in this project. The reading and processing of the data and the algorithms development were made with Python programming language, using Jupyter Notebook as a programming environment, because it supports all libraries required, it contains live code, equations and visualisations. The results were analysed using QGIS which was connected to a spatial database on PostegreSQL and the statistical results were interpreted on Microsoft Excel. This is the brief description of the technologies used:

- Python: Free and open source, programming language that is equipped with multiple packages and libraries of data mining and machine learning. It also has an automatic management of the memory and dynamic data structures;
- PostgreSQL: Free and open source object-relational database designed for data administrators and developers;

- QGIS: Free and open source Geographic Information System (GIS), which allows to create, edit, visualise, analyse and publish geospatial information.

There are a few Python packages that were used in the processing of the data and the algorithm's development:

- Pandas: Open source library offering high-performance, easy-to-use data structures and data analysis tools;
- Numpy: Regards to scientific computing and was used to conduct multiple calculations with location data and distances;
- Matplotlib: Library for generating high quality graphs and plots;
- Geopy: Python client for several popular geocoding web services that facilitates the coordinates location;
- kneed: Knee-point detection, used to analyse graphs;
- Scikit-learn (sklearn): A set of python modules for machine learning and data mining used, in this project, to apply clustering algorithms.

This page is intentionally left blank.

Chapter 5

Experimental Results

In this chapter, the results obtained after implementing the methodology described in Chapter 4 are presented. It exhibits the outcomes from the customer segmentation and the characterization of each group that emerged from it, the outcomes from the important places identified, and the results from the validation and evaluation process. The last section is dedicated to the scalability of the algorithm to identify important places.

5.1 Customer Segmentation and Sleeping Periods

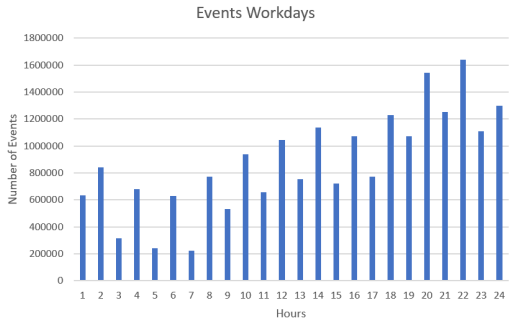
Before starting implementing the algorithm to segment the clients, the base of the clustering must be well defined. So, we made a study on the frequency of events, of the user, throughout the different days of the week.

A study made by Abdullah et al. [1] investigated the wake cycle of nine individuals, finding that most of them varied their sleep time patterns between workdays (Monday to Friday) and weekends (Saturday and Sunday) and changes in their mid-sleep time during different seasons of the year. Based on that, since June, July, and August are considered vacation seasons, these months were separated and a study on the number of events throughout the weekdays from September to October was made.

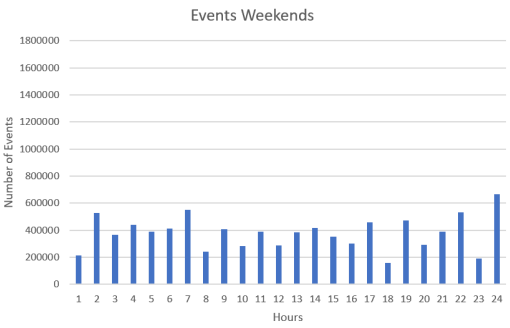
Figure 5.1 gives a visualization of time periods when the users are most and less active during the week and proves that routines vary between the different days of the week. The graph illustrated in 5.1(a) shows the frequency of events on the hours of the day during workdays and illustration 5.1(b) the frequency on weekends.

It is observable that the number of events from Monday to Friday is slightly higher than on weekends, which might be because these weekdays are typically the workdays, and weekends are most of the time leisure time. It is also notable that the workdays have less activity early in the morning and it starts growing around 8AM, which is around the usual commuting to work happens.

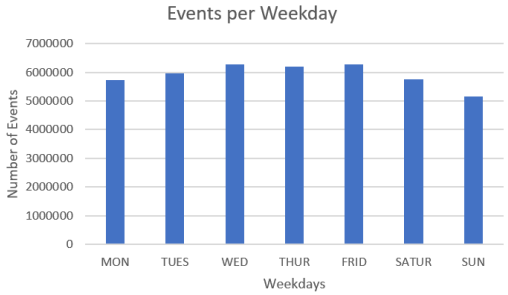
This analysis meets the results obtained by Abdullah et al. [1] and proves that the 30 729 subscribers on the dataset, have different routines during the different days of the week. So, an analysis of the routines and several experiments, taking into account the typical working hours in Portugal, were performed. Finally, we concluded that the best scenario, to group subscribers, was based on the events made during the night (12AM-4AM), the morning (5AM-10AM), and the afternoon (1PM-5PM) at workdays.



(a) Events per workdays



(b) Events per weekends



(c) Events per weekdays

Figure 5.1: Event frequencies for the entire dataset

As mentioned previously, the elbow/knee method was used to obtain the optimal k value to apply in K-Means. Then, the value returned was used to identify k groups of clients. The clients in each group must have similar behaviors, based on their activity on the mobile network throughout the identified slots of the day.

After analyzing the mean number of events that each user has from 12AM to 4AM, 5AM to 10AM, and 1PM to 5PM, which we considered, respectively, the night hours, the morning hours, and the afternoon hours, the elbow/knee method returned an optimal $k = 3$ (figure 5.2).

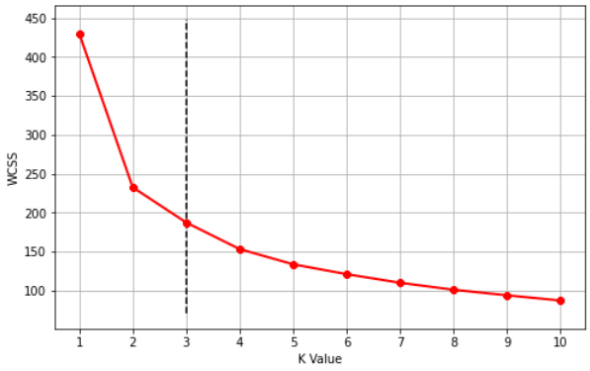
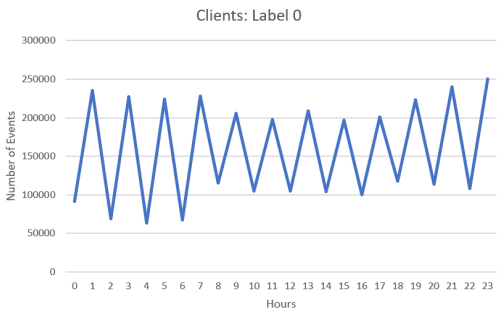


Figure 5.2: Elbow/knee method to determine k

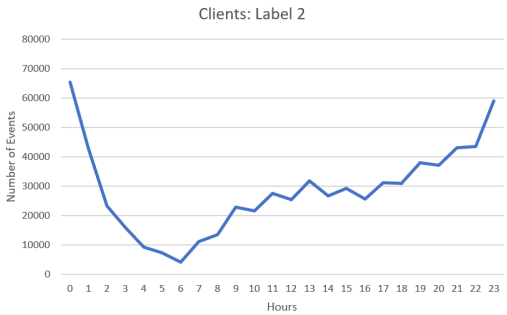
With the number of clusters defined, K-Means was applied to perform the segmentation of the subscribers. Then, three different groups with distinct behavior characteristics, were spotted.

As presented in figure 5.3, the groups identified are very distinct from each other. Group 1 comprises 13 703 users (44,59%) from the final dataset and was labeled with label 0.



(a) Group 1

(b) Group 2



(c) Group 3

Figure 5.3: Customer segmentation

The users of this group were determined as "undetermined workers" because, in the graph (figure 5.3(a)), the sleeping period is not clearly visible and has to be inferred. The second group, Group 2, presented by figure 5.3(b) was labeled with label 1 and comprises 14 146 users (46,03%) that, due to the reduced activity during the night, were determined as "day workers". Finally, Group 3 (figure 3(c)), which contains 2 880 users (9,38%) identified as "night workers", due to their high activity during the night hours, was labeled with label 2.

Based on the determination of the type of users in each group, a slot that we assume to be part of the users sleeping period, was determined. Then, considering the analysis made and the Portuguese work system, the working hours were also determined. Since the sleeping period of Group 1 is not directly visible, based on the common sleeping hours in Portugal, the sleeping period of this group was determined to be from 12AM to 4AM and the working hours from 1PM to 5PM in a way that embraces users that leave their work at 4PM and the ones that start working at 4PM [16]. Group 2 was determined to be constituted by "day workers" based on their activity, so, the sleeping period was established between 2AM and 5AM, and work hours from 9AM to 1PM, based on the typical Portuguese working hours. For the users that were identified as "night workers" (Group 3), the sleeping period was determined to be from 5AM to 9AM, and their working hours from 12AM to 4AM. This last slot of working hours was decided according to the Portuguese work system, which establishes that for a night worker, the working hours are some period of time between 10PM and 7AM of the following day [9].

The label associates the user to a profile that ultimately is going to determine which criteria with time constraints are going to be applied to the user's events to determine home and work locations. Table 5.1 shows a sample of the final dataset labeled.

The variables used to identify the optimal k value were only found after several experiments. At the beginning of the experiments, we were using a slot from 7PM to 11PM, that we

consider being the evening events. However, using also this slot, a $k = 4$ was returned, and after performing the cluster, four groups were identified, one similar to the one illustrated in figure 5.3(a), and the other three with similar night routines between them. Therefore, we concluded that when using only three slots as metrics, we can clearly extract the different routines of each group.

Table 5.1: Sample of the final dataset labeled

UserID	CellID	...	TYPE	Label
6D38B...	30439	...	CDR-WEEKDAY	2
EC7A7...	1946409	...	SNAPSHOT-WEEKDAY	1
4BD80...	609804	...	SNAPSHOT-WEEKEND	1

5.2 Geo-Profiling: Home, Second Home, and Work

As explained, VDBSCAN was used in this project to geo-profile users, by determining their important places. The model treats user by user by filtering his/her events and clustering them according to the profile that is associated with the label attributed. It is also in this phase that the events associated with unknown antennas (latitude and longitude with negative values) are ignored/excluded.

Figure 5.4 shows how the model identifies the important places. First, it starts by returning the user's label and the profile associated. Next, it gives the antennas found to represent home, second home, and work locations and indicates the *eps* value used to perform the clustering in each situation.

```
USER: 299FBB6A7D1F33013DC9113AB269311617D344E7F3DE5B8703DDD017385FDBFD
Label 1: Day worker
SP 2AM - 6AM & work hours 9AM - 1PM
Clustered 6 points into 1 clusters, with 83.33% compression in 0.032 seconds for HOME location
EPS: 0.010000 degrees (~1110.0 meters)
   cellid  VAL_LATITUDE  VAL_LONGITUDE  NumEvents
0  893196      40.17692       -8.45021          6
                        ***************************
Clustered 266 points into 1 clusters, with 99.62% compression in 0.059 seconds for SECOND HOME location
EPS: 0.009556 degrees (~1060.74845968558 meters)
   cellid  VAL_LATITUDE  VAL_LONGITUDE  NumEvents
0  572683     40.241217      -8.445423        266
                        ***************************
Clustered 86 points into 1 clusters, with 98.84% compression in 0.105 seconds for WORK location
EPS: 0.010000 degrees (~1110.0 meters)
    cellid  VAL_LATITUDE  VAL_LONGITUDE  NumEvents
0  1499435     40.263269      -8.429343         86
```

Figure 5.4: User Example

To have a better perspective on the obtained results, we can visualize them on QGIS. Figure 5.5 shows all events of user that was presented in figure 5.4. Each point on the map layer represents an antenna where one or more events of the user were registered from July to October of 2020. It is also observable the number of events that were registered in each antenna.

Figure 5.6 represents the events on each matrix and the point that is marked represents the antenna that was identified as home (a), second home (b), and work (c) locations, respectively.

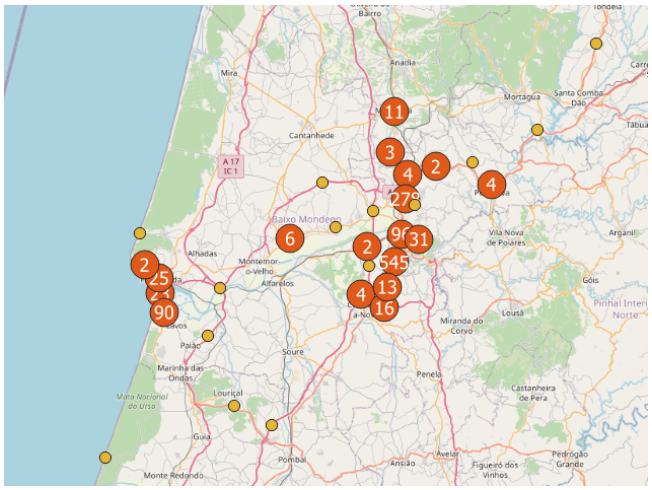
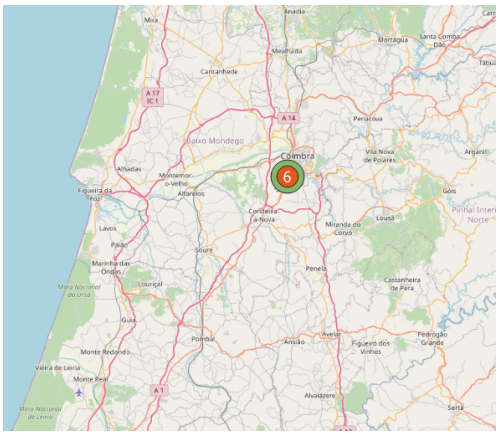
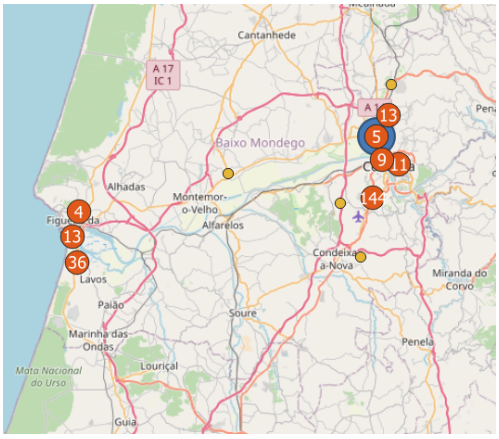


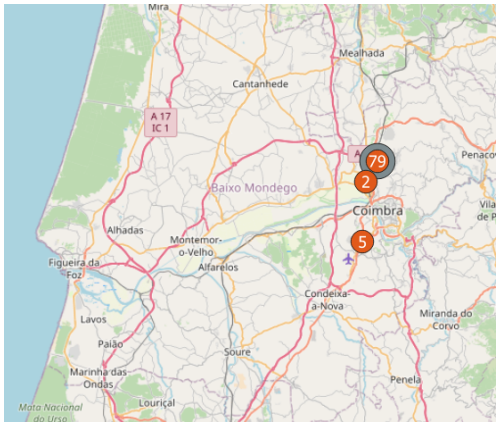
Figure 5.5: All events of the user



(a) Home



(b) Second Home



(c) Work

Figure 5.6: Home, Second Home, and Work locations of the user

This process is equal for all users in the dataset. There are situations in which second home locations are identified in the same antennas that were determined as home locations. In these situations, the algorithm informs that the user does not have a second home location (figure 5.7).

In other cases in which more than one location is identified as the place that we are looking for, the column with the number of events in each cluster ("NumEvents") can be helpful

```
USER: 00F368B6180A6F8F87F8F59BCE2234D73194B01FCA10159BF5F7F7C49B529E6A
Label 0: Indetermid
SP 12PM - 4AM & work hours 1PM - 5PM
Clustered 94 points into 1 clusters, with 98.94% compression in 0.041 seconds for HOME location
   cellid  VAL_LATITUDE  VAL_LONGITUDE  NumEvents
0   40152     40.188067      -8.448094         94
                        **************************
Clustered 781 points into 1 clusters, with 99.87% compression in 0.137 seconds for SECOND HOME location
EPS: 0.009442 degrees (~1048.0992422950771 meters)

NO SECOND HOME: The second home location in the first home location

                        **************************
Clustered 236 points into 1 clusters, with 99.58% compression in 0.174 seconds for WORK location
EPS: 0.009442 degrees (~1048.0992422950771 meters)
   cellid  VAL_LATITUDE  VAL_LONGITUDE  NumEvents
0   30395      40.19997        -8.4318        235
```

Figure 5.7: No Second Home

to determine which location has more probability of being the one that we are looking for: home locations are the locations with less activity, since behind this search is a period with less activity, and, contrary to that, work locations are the location with more events because we are looking for the antenna most regularly used in terms of days and hour slot. Figure 5.8 shows an example of a user with two identified workplaces. In this case, we assume that it is more likely that the second antenna (CellID: 8005) represents the workplace.

```
USER: A9195C97D496BBD45593135AC85AAF64FB2821723BB78D5D2B971C292CFCA852
Label 0: Indetermid
SP 12PM - 4AM & work hours 1PM - 5PM
Clustered 95 points into 1 clusters, with 98.95% compression in 0.027 seconds for HOME location
EPS: 0.112353 degrees (~12471.129332051887 meters)
   cellid  VAL_LATITUDE  VAL_LONGITUDE  NumEvents
0    8005     39.817381      -8.108378         95
                        **************************
Clustered 478 points into 1 clusters, with 99.79% compression in 0.075 seconds for SECOND HOME location
EPS: 0.083150 degrees (~9229.603253730042 meters)
   cellid  VAL_LATITUDE  VAL_LONGITUDE  NumEvents
0   22485      39.81379       -8.19145        478
                        **************************
Clustered 102 points into 2 clusters, with 98.04% compression in 0.118 seconds for WORK location
EPS: 0.003695 degrees (~410.1962310357887 meters)
   cellid  VAL_LATITUDE  VAL_LONGITUDE  NumEvents
0   21217     40.217160      -8.411299         13
1    8005     39.817381      -8.108378         89
```

Figure 5.8: More than one workplace

The model puts the results in a csv file to latter be analysed. Table 5.2 shows how the locations are placed in the outcomes file.

Table 5.2: Outcome file

UserID	CellID	LAT	LONG	Label	PLACE
299FB...	893196	40.17692	-8.4502	1	HOME
299FB...	572683	40.241217	-8.445423	1	SEC HOME
299FB...	1499435	40.263269	-8.429343	1	WORK

5.3 Validation and Evaluation

As mentioned, in this project, we only have available ground-truth to validate and evaluate the results from the inference of the home locations. This process was made based on the

information we have available on the datasets. As explained, first the antennas of Coimbra's district were analyzed, and after we observe the densities in more dense (urban areas) and less dense (rural areas) areas, we analyzed the average radius of the antennas in each area, and we use that information to evaluate the home results.

Using the density-based DBSCAN two clusters were identified, enhancing the high density of antennas in the cities of Coimbra and Figueira da Foz (urban areas). All the other regions were considered zones of low density of antennas (rural areas), as illustrated in figure 5.9.

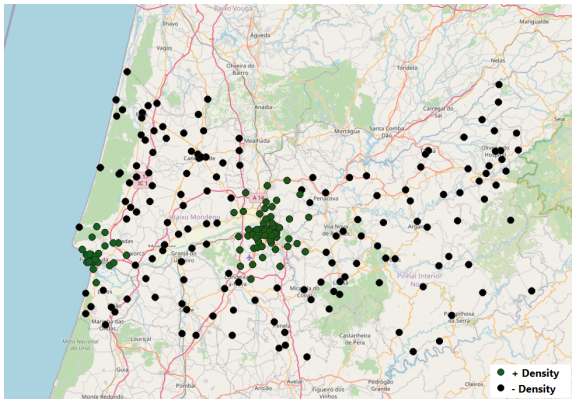


Figure 5.9: Density zones in Coimbra (Antennas)

After performing the distinction, the antennas in the two zones were analyzed. Table 5.3 presents the analysis executed, indicating that the mean coverage radius of the antennas in high dense areas is approximately 1 900m and in less dense areas is 2 820m. Since in urban zones, the coverage area is smaller, it was expected a more precisely identification of the place in this zones.

Table 5.3: Sample of CDRs

	+Density	-Density
Antennas	942	5246
Min Radius	68m	73m
Max Radius	10 000m	23 000m
Mean Radius	1 900m	2 820m

Next, we determined the home location of the users on the ground-truth dataset. Since only the antennas in the district of Coimbra were analyzed, to make the process consistent, only the users in which the home location was found in the antennas of Coimbra's district were taken into account. Thus, a final sample of 3 838 users was used to evaluate the dataset.

After, we measure the straight line distance between the antenna found to represent home location and the centroid. As explained, distances below the thresholds were considered scenarios of success.

After this process, we realized that some antennas identified as the home location were very distant from the real home locations and some users that presented these situations, had the majority of their events registered in Coimbra and none or just a few in their real home area. So, we recognize these users as clients that live in Coimbra, but their home location is registered out of Coimbra (e.g., students). To perform a correct evaluation, these situations were considered annotation errors.

In order to understand from which value the distances should be considered annotation errors, we executed three experiments: in the first experiment, the annotation errors were considered situations where the distance between the two locations was higher than 60km, in the second experiment the distance was higher than 40km, and in the third higher than 20 km. After performing the experiments, we achieved 61% of accuracy for the first experiment, 62% for the second, and 66% for the third.

The distances were measured in a straight line, but this distance does not correspond to the real road trip between the two places. So, we considered that 20km measured in a straight line was a reasonable value, and distances higher than this were treated as annotation, which means that users in this situation were considered residents in Coimbra, but their billing address is far away from the place that was inferred as home location. The 780 locations found, that presented this scenario, were discarded from the evaluation. Table 5.4 presents the results of the evaluation, after discarding the annotation errors.

Table 5.4: Evaluation for accuracy

Area	Home locations found	Distance $<$ Threshold	Accuracy
+Dense	2 706	1 830	68%
-Dense	406	209	51%
Results	3 112	2 039	**66%**

As expected the accuracy in urban areas is higher, which is justified by the higher density of antennas with lower coverage area. These factors increase the precision of the identification [25].

Once the evaluation was performed, we analyzed a sample of 200 users, from the 3 838 on the evaluation dataset, to understand the types of errors/situations. The sample was composed of 44% of users with label 0, 45% of users with label 1, and 11% of users with label 2, which are the approximate percentages of the labels in the final dataset. All of these users were selected randomly. The table presented in figure 5.10 presents the types of situations identified and their description.

Situation	Note	Description
Type 0	**Annotation errors**	Outliers - 35 (17,5% of the 200 homes identified)
Type 1 **113 (68,5%)**	**Success**	The antenna identified is near the real home location (Distance < 1 900m/2 820m)
Type 2 **6 (3,6%)**		The antenna identified is the closest from the real home, but the distance is bigger than the established
Type 3 **14 (8,5%)**		The antenna identified to represent the home location is not the nearest, but it is in second line (sometimes due to the larger coverage of the antennas)
Type 4 **32 (19,4%)**	**Unsuccess**	The SP is properly determined but the home location is not correctly determined

Figure 5.10: Typification of scenarios

With the analysis of the sample, four types of situations were identified. The first one is the situation of success when the real home is at less than 1 900m or 2 820m, in the respective areas. Then, we found two types of situations that we considered as being flaws caused by the large coverage of some antennas. These types of situations occurred especially (1) in high dense zones when the antenna that was found to represent home is not the one nearest

to the centroid but is in the area around, and (2) in the periphery of high dense zones and less dense zones, when the distance between the centroid and the nearest antenna, which is also the antenna identified as home, is higher than the established. The other type of situation is when the sleeping period was properly determined, because it was found the slot with less activity, yet the distance between the two places is elevated.

Based on the analysis of the sample, rules were defined to identify the types of situations on the evaluation dataset. Through these rules, it was established that: until 1 900m or 2 820m, depending on the zone, the home locations were accurately identified, between these values and 3 500m, we can still affirm that we are in the area of the user's home, however, with a lower degree of certainty, and that more than this distance the home location was not well determined.

Situation	Note	Description
Type 0	**Annotation errors (Outliers)**	Distance > 20 000m
Type 1	**Success**	+Density: Distance < 1 900m -Density: Distance < 2 820m
Type 2		+Density: Distance > 1 900m & Distance < 3 500m -Density: Distance > 2 820m & Distance < 3 500m
Type 3	**Unsuccess**	Distance > 3 500m & Distance < 20 000m

Figure 5.11: Rules to identify different scenarios

The rules were then applied to the evaluation results and for home location identified, we were able of identifying the type of situation. Figure 5.12 shows that the results on the sample were representative.

Situation	Note	Description
Type 0	**Annotation errors**	Outliers - 780 (20% of the 3 892 homes identified)
Type 1 **2039 (66%)**	**Success**	The antenna identified is near the real home location (Distance < 1 900m/2 820m)
Type 2 **482 (15%)**		The antenna identified is the closest from the real home, but the distance is bigger than the established
		The antenna identified to represent the home location is not the nearest, but it is in second line (sometimes due to the larger coverage of the antennas)
Type 3 **600 (19%)**	**Unsuccess**	The SP is properly determined but the home location is not correctly determined

Figure 5.12: Scenarios/Situations identified

5.4 Scalability

VDBSCAN is tuning version of DBSCAN that before adopting the traditional DBSCAN selects suitable *eps* parameters for different densities. Thus, it is suitable to affirm that the base of this model is DBSCAN. This algorithm has an average run-time complexity of $O(n \log n)$ in the best scenario and in a worst-case scenario, it comes to $O(n^2)$.

We studied the run-time complexity of the algorithm that in this model identifies the

important places. The graph obtained is presented in figure 5.13.

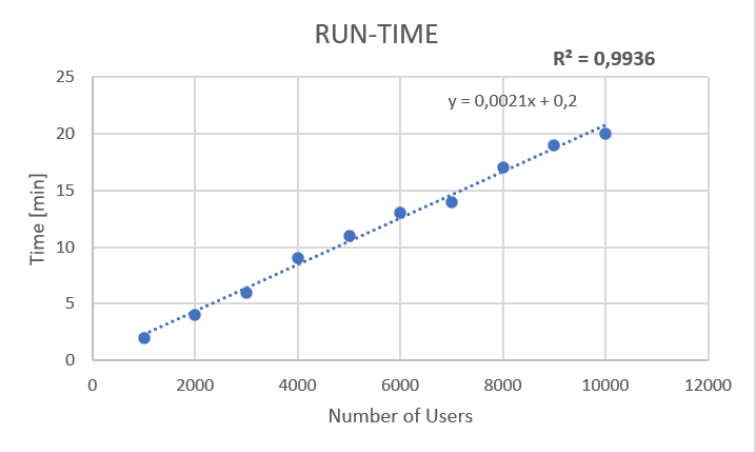


Figure 5.13: Run-Time complexity analysis

To perform this analysis we observe the average running time of the algorithm when the home, second home, and work of 1 000 users are determined. Then the experiment was repeated from 1 000 in 1 000 until 10 000 users. The ten experiences were performed five times each, and the average time of each experience is presented in the graph.

By analyzing figure 5.13 and the equation of the line presented in the graph, we conclude that the run-time complexity of the algorithm is linear, which tells us that we are not facing the worst scenario with a complexity of $O(n^2)$.

We also need to be aware that this experience was performed on a personal computer and that in this phase, the model treats user by user, so, this process can be parallelizable.

With this analysis, we conclude that, in a personal computer, the algorithm took an average milliseconds to identify the important places of one user, and about 20 minutes to identify the home, second home, and work locations of 10 000 users.

Chapter 6

Conclusion

While location data, as CDRs, alone means little for the general public, its use in location analytics has the power to transform these raw and, sometimes, irrelevant data into meaningful insights.

In this thesis we addressed the inference of home, second home, and work locations from CDRs, concluding that this data offers a great and viable way to infer important places without compromising the users' time and privacy since we worked with ubiquitous and anonymized data.

This project was developed under in an internship at AmILab, started in September of 2020, counted with the collaboration of a telecommunication company, and shows that for today's changing world, the power of "where" is relevant for countless applications across industries and organizations.

This chapter addresses the project's development steps, the main contributions produced, the challenges encountered, and highlights a few challenges that deserve to be addressed in future work.

6.1 Development of the Project

The project was divided into several phases, being the first one the research on the state-of-the-art. While researching, we began analyzing the data to understand its characteristics and start processing it. Both CDRs and Snapshots dataset had differences, so many pre-processing phases had to be executed before we start inferring on important places.

Then, we started developing a model with temporal assumptions, that treated all the population on the dataset the same way. In this initial model, the temporal criteria were based on the work of Isaacman et al. [16] and consisted of two time windows to identify home and work locations of the users: home locations was the place where events were registered, throughout the week from 7PM to 7AM and work locations where events were registered on workdays from 1PM to 5PM. After filtering the events, DBSCAN was applied. The selection of the algorithm was made based on the work of Yang et al. [39], which used it to eliminate the noise caused by the load sharing between antennas and identify home and work locations.

Once we started obtaining results, we quickly realized that VDBSCAN could be superior and tuned the algorithm. We use the method described on section 4.1.3 to validate and

evaluate this approach. We obtained a 71% of accuracy, which is better compared to the 69.02% of accuracy obtained by Lumpsum et al. [34] in their work "Exploring Home and Work Locations in a City from Mobile Phone Data". The accuracy obtained using this model is inferior to the 88% obtained by Isaacman et al. [16], however in this work they evaluated the results of only 19 volunteers.

Despite reasonable results achieved with the initial model, one objective of the project was the determination of each client profile, taking into account the registers/activities on the mobile network. For that, we had to discard the criteria applied and return to the state-of-the-art. We started by studying several techniques to identify sleeping periods to replace the criteria used to identify the home and work locations.

We ended up implementing three techniques to identify a period in which it is most probable that the user is at home, the sleeping period. First, we used the technique of Lumpsum et al. [34] and for each user, we identified an hour with less activity to determine the sleeping hour and encounter the sleeping period based on that. This technique resulted in a sleeping period of nine hours: the sleeping hour and the four hours before and after. Next, we proceeded with an experiment that identified a five hour slot with minimum activity for each user, resulting in a five hour sleeping period for each user. Finally, we used the client segmentation to group clients based on their activities during the hours of the day and then profile them according to the slots with less activity. This last technique resulted in a five hour sleeping period, based on the analysis of the behavior of the users in each group.

After implementing these three techniques, they were integrated with VDBSCAN to identify home locations. The evaluation resulted in a 60% of accuracy for the first technique, 64% for the second one, and 66% for the one involving the client segmentation, being the last one the selected to be part of the methodology.

Although the 71% of accuracy obtained with the methodology implemented in the first phase was better than the 66% obtained with the adopted, the profile/characterization of subscribers was a requirement of the telecom company and goal that we expected to achieve since the beginning of the project. The most probable reason for these results is the size of the slot to identify the home location. In the first model we used a 12 hour slot to identify home locations and in the second model, we reduced the slot to five hours, because of the overlapping slots between different groups.

Hence, these accuracy values were discussed in the meetings held between the AmILab team and the telecom company, and it was established that it was preferable to achieve a lower percentage of accuracy rather than to apply fixed criteria to the entire dataset and not be able to identify the clients routines.

The segmentation process was a process that required multiple experiments. While establishing the variables to the base of the segmentation, we started by using only CDRs, however, the segmentation using only this type of data did not allowed us to identify clients with clear distinct routines. Only with the use of Snapshots, this segmentation was possible.

The validation and evaluation processes were discussed by both teams, and due to the data available, their experience and the late stage of the process, the method described previously (Section 4.1.3) was followed. This method was also used by Mamei et al. [25] that used to mean coverage radius of the antennas to evaluate the results.

After achieving the results mentioned above, we analyzed if, with the usage of the median of the coverage radius of the antennas to perform the evaluation, we could achieve better

results. However, for both areas, urban and rural, the median was smaller than the mean: 1 212m to high dense areas and 2 422m for less dense areas. When using the mean coverage, we faced some situations where the distance between the centroid of the postal-code of the real home local and the antenna found, were bigger than the established by the mean, then, if we used a smaller value, like the median, these situations would be more frequent.

After exploring the mean and the median values of the coverage radius, we concluded that following the method of Mamei et al. [25] and using the mean coverage radius for both urban and rural areas was an asset to our results.

Then, we started studying the district of Coimbra in terms of area. This district has an area of 319.4Km^2. Next, we calculated the square root of this area, which is 60km, and decided that distances between real home location and the inferred home location, measured in a straight line, higher than this value, were annotation errors provoked by users that had their home place registered out of Coimbra, but were living in Coimbra at that time (September and October of 2020). However, after a few experiments, we concluded that this was a very long distance between two places if a road trip was taken, so, instead of distances higher than 60km, we determined that distances higher than 20Km also in a straight line were discarded.

6.2 Main Contributions

At an early stage of this work, when reviewing the state-of-the-art to identify meaningful places, we targeted the outcomes that we intend to achieve. This exploratory work also allowed us to establish our goals, strengths, and main challenges related to the data sources. Comparing this work with some published articles, this one is scalable (supports high volumes of data) and determines important places based on the user profile (night or day workers) to build his/her geographic map.

Besides these main contributions, others were achieved, such as the use of the new type of data, the Snapshots, that presented features essential to perform the segmentation. We also made an analysis on the sleeping periods throughout different days of the week and concluded that individuals vary their routines during the week. Adding to that, we implemented several techniques to determine sleeping periods, and evaluate the results provided by the model using each technique. Although only one technique was applied, the others can be suitable for future projects.

Before starting the evaluation and validation processes, the telecom company informed us that evaluations performed on other models developed by their teams, showed an average of 30% of accuracy and that these results were achieved mostly because some antennas cover entire cities. Nonetheless, it is important to understand that, although the model developed in this project, compared to the others is not achieving higher accuracy [16] [34], the 66% of accuracy obtained is twice than the achieved by the telecom company until now.

Moreover, the results on the home locations identified by the model were validated and evaluated with real and updated data, and using a method that was also used by other authors [25].

Ultimately, this work proves that besides the controversial opinions on using CDRs to determine important places, this type of data allowed us to identify home locations with 66% of accuracy, taking also into account that 15% were considered flaws caused by the large coverage area of some antennas.

6.3 Challenges

Since the beginning of the project, we were aware that the telecom company has to face a series of approvals to provide us data. A few weeks after starting the project, we had only Snapshots, so we started developing the project with only this data. Then, CDRs were provided and we had to adapt the model and all the processes.

Since main goal was the inference of important places from CDRs, we tried to develop an approach with only this type of data. However, after multiple experiments we conclude that when using only this data, the segmentation did not allow us to group users with clear distinct routines, and we had to adopt also the snapshots data.

Another challenge was the instability of the World due to the pandemic situation that limits the general movement of the population, especially the displacement to the workplace. Due to this situation, work locations are frequently identified in the same location as home, since a significant part of the population was, at that time, working from home. Though, this does not invalidate the model, because we are still identifying work locations.

The ground-truth to perform the home locations validations was provided later in the course of the project, which means that some tasks in our schedule had to be re-arranged. With this, the method to perform it had to be adapted according to the ground-truth dataset and the project that was already developed.

To validate the inference of second homes and work locations we proposed to the telecom company the realization of a survey, however, this type of ground-truth involves challenges, especially when it comprises personal aspects. Although they are still trying to negotiate with the company to provide us this data for future work, the data was not available for this thesis.

6.4 Future Work

For future work, there are a few points that deserve to be highlighted. It is important to recap that one group resultant from the customer segmentation (section 5.1), named "group 1" and labeled with label 0, was identified as "undetermined workers" because it was impossible to have a clear vision of the clients routines. We assumed that this group of clients, works during the day and their sleeping period is at night, yet this group can be studied in deep to try to understand their real profile.

After having a clear vision of the profiles, the next step would be to obtain ground-truth no validate the work and the second home locations. Next, we could use the evaluation process, to understand how the size of time slot affects the accuracy of the inference.

The ground-truth for work and second homes and the more intense study of the profiles will allow an improvement on the performance of the model. Then, we can use it with data from other cities/clients and understand the results obtained by the telecom company on the evaluation of their models.

With the geo-profile validated, this project can be integrated into another one that is being developed at AmILab and that consist of inferring the check-ins and check-outs of the users from their important places. Then, based on that we could infer commuting trips and use that to various activities, such as transport planning or environmental decisions.

The work developed in this thesis will be the basis of the writing of a scientific article.

This article is going to present the model developed, to geo-profiles users, through the identification of their home, second home, and work locations using CDRs and without fixed assumptions on the entire set of users.

This page is intentionally left blank.

References

[1] S. Abdullah, M. Matthews, E. L. Murnane, G. Gay, and T. Choudhury. Towards circadian computing: "early to bed and early to rise" makes some of us unhealthy and sleep deprived. 2014.

[2] R. Ahas, S. Silm, O. Järv, and E. Saluveer. Using mobile positioning data to model locations meaningful to users of mobile phones. 2010.

[3] D. Arthur and S. Vassilvitskii. k-means++: the advantages of careful seeding. 2007.

[4] BASILB2S. In-depth intuition of k-means clustering algorithm in machine learning. URL `https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/`.

[5] M. Berlingerio, F. Calabrese, G. D. Lorenzo, and X. Dong. Safercity: A system for detecting and analyzing incidents from social media. 2013.

[6] O. Burkhard, R. Ahas, E. Saluveer, and R. Weibel. Extracting regular mobility patterns from sparse cdr data without a priori assumptions. 2017.

[7] D. Cao and B. Yang. An improved k-medoids clustering algorithm. 2010.

[8] C. Chen, L. Bian, and J. Ma. From traces to trajectories: How well can we guess activity locations from mobile phone traces? 2013.

[9] D. da República Eletrónico. Código do trabalho. URL `https://dre.pt/web/guest/legislacao-consolidada/-/lc/108165886/201710020500/73481712/diploma/indice/12`.

[10] I. Dabbura. K-means clustering: Algorithm, applications, evaluation methods, and drawbacks. URL `https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644`

[11] D. Dai and T. J. Oyana. Association rules and frequent item sets. 2006.

[12] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. 1996.

[13] B. Furletti, L. Gabrielli, C. Renso, and S. Rinzivillo. Identifying users profiles from mobile calls habits. 2012.

[14] M. C. González, C. A. Hidalgo, and A.-L. Barabási. Understanding individual human mobility patterns. 2008.

[15] Hahsler, Michael, B. Grun, and K. Hornik. Association rules and frequent item sets. In *Journal of Statistical Software*, 2005.

[16] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, and J. Rowland. Identifying important places in people's lives from cellular network data. 2011.

[17] S. Isaacman, R. B. R. Cáceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger. Human mobility modeling at metropolitan scales. 2012.

[18] S. Jiang, J. Ferreira, and M. C. Gonzalez. Activity-based human mobility patterns inferred from mobile phone data: A case study of singapore. 2017.

[19] G. Karypis, E.-H. S. Han, and V. Kumar. Chameleon: A hierarchical clustering algorithm using dynamic modeling. 1999.

[20] R. Kujala, T. Aledavood, and J. Saramäki. Estimation and monitoring of city-to-city travel times using call detail records. 2016.

[21] P. Lathiya and R. Rani. Improved cure clustering for big data using hadoop and mapreduce. 2016.

[22] A. Lind, A. Hadachi, P. Piksarv, and O. Batrashev. Spatio-temporal mobility analysis for community detection in the mobile networks using cdr data. 2017.

[23] P. Liu and D. Z. andNaijun Wu. Vdbscan: Varied density based spatial clustering of applications with noise. 2007.

[24] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson. Approaching the limit of predictability in human mobility. 2013.

[25] M. Mamei, M. Colonna, and M. Galassi. Automatic identification of relevant places from cellular network data. 2015.

[26] B. Mihai-Florin, R. Adrian, and M. I. Liviu. Prepaid telecom customers segmentation using the k-mean algorithm. 2012.

[27] A. Nagpal, A. Jatain, and D. Gaur. Review based on data clustering algorithms. 2013.

[28] E. Nandapala and K. Jayasena. The practical approach in customers segmentation by using the k-means algorithm. 2020.

[29] T.-A. Nguyen. Customer segmentation: A step-by-step guide for growth. URL `https://openviewpartners.com/blog/customer-segmentation/#.YKqpaKhKg2w`.

[30] J. Qi, Y. Yu, L. Wang, and J. Liu. K*-means: An effective and efficient k-means clustering algorithm. 2014.

[31] G. Ranjan, H. Zang, Z.-L. Zhang, and J. Bolot. Are call detail records biased for sampling human mobility? 2012.

[32] A. Saarik. Trajectory reconstruction and mobility pattern analysis based on call detail record data. 2017.

[33] R. Shibasaki. Call detail records (cdr) analysis: Republic of guinea. 2017.

[34] L. Tongsinoot and V. Muangsin. Exploring home and work locations in a city from mobile phone data. In *2017 IEEE 19th International Conference on High Performance Computing and Communications, IEEE 15th International Conference on Smart City, IEEE 3rd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, 2017.

[35] M. Vanhoff, F. Reis, T. Ploetz, and Z. Smoreda. Assessing the quality of home detection from mobile phone data for official statistics. 2018.

[36] M. Vanhoff, F. Reis, T. Ploetzand, and Z. Smoreda. Detecting home locations from cdr data: introducing spatial uncertainty to the state-of-the-art. 2018.

[37] H.-Y. Wan, Y.-F. Lin, Z.-H. Wu, and H.-K. Huang. Discovering typed communities in mobile social networks. 2012.

[38] J. Xi. Spatial clustering algorithms and quality assessment. 2009.

[39] P. Yang, X. Wan, T. Zhu, and X. Wang. Identifying significant places using multi-day call detail records. 2014.

[40] S. Yildirim. Customer segmentation: A step-by-step guide for growth. URL `https://towardsdatascience.com/dbscan-clustering-explained-97556a2ad556`.

[41] Z. Zhao, S.-L. Shaw, Y. Xu, F. Lu, J. Chen, and L. Yin. Understanding the bias of call detail records in human mobility research. 2016.

This page is intentionally left blank.

Appendices

This page is intentionally left blank.

Appendix A: Gantt Charts

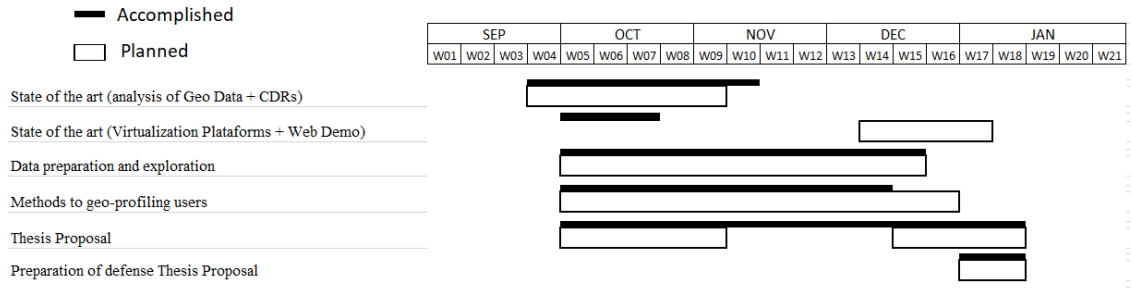


Figure 1: Gantt Chart of the first semester - planned *vs* accomplished

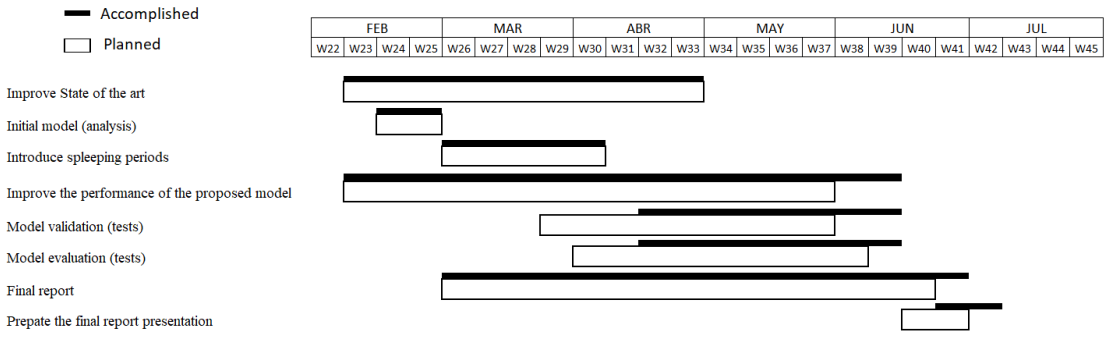


Figure 2: Gantt Chart of the second semester - planned *vs* accomplished

This page is intentionally left blank.

Appendix B

Algorithm 1 Home, Workplace, and Second Home

```
for user = 1, 2, ... do
    Read label
    if label=0 then
        HomeDataset = Sept and Oct workweek SP from 12PM to 4AM
        WorkDataset = Sept and Oct workweek WH from 1PM to 5PM
        SecHomeDataset = Jul and Aug all days and Sept and Oct weekends from
8PM to 8AM
    else if label=1 then
        HomeDataset = Sept and Oct workweek SP from 2AM to 6AM
        WorkDataset = Sept and Oct workweek WH from 9AM to 1PM
        SecHomeDataset = Jul and Aug all days and Sept and Oct weekends from
8PM to 8AM
    else
        HomeDataset = Sept and Oct workweek SP from 5AM to 9AM
        WorkDataset = Sept and Oct workweek WH from 11PM to 4AM
        SecHomeDataset = Jul and Aug all days and Sept and Oct weekends from
8PM to 8AM

    if len(HomeDataset)=0 then
        if label=0 then
            newHomeDataset = Sept and Oct workweek SP from 10PM to 6AM
        else if label=1 then
            newHomeDataset = Sept and Oct workweek SP from 12PM to 8AM
        else
            newHomeDataset = Sept and Oct workweek SP from 3AM to 11AM

        if len(newHomeDataset)=0 then Impossible to determine Home Location
        else
            vdbscan (newHomeDataset)                          ▷ Home location

    else
        vdbscan(HomeDataset)                                  ▷ Home location

    if len(WorkDataset)<5 then Impossible to determine Work Location
    else
        vdbscan (WorkDataset)                                 ▷ Work location

    if len(SecHomeDataset)<5 then Impossible to determine Second Home Location
    else
        vdbscan (SecHomeDataset)                              ▷ Second Home location
```

Figure 3: Algorithm to identify Home, Second Home, and Work locations