NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Urban Mobility Patterns Based on Mobile Phone Data

Clustering urban zones in Lisbon

Lucas Emilio Mendes Ferreira

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

URBAN MOBILITY PATTERNS BASED ON MOBILE PHONE DATA

Clustering urban zones in Lisbon

by

Lucas Emilio Mendes Ferreira

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science

Supervised by

Supervisor: Miguel de Castro Neto, PhD, NOVA Information Management School

Co-Supervisor: Bruno Jardim, PhD, NOVA Information Management School

July, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, July 2024

DEDICATION

First, I want to dedicate my thesis to my parents, Cintia and Gustavo, for always supporting me since I was a child. It was because of you that I learned how to study and always aimed to improve myself. Living with you and my sister Leticia helped me to grow as a person and be confident that I will always have a safe place with my family.

In addition to them, my fiancée, Carolina, was the one who encouraged me to pursue a master's degree and this career path. Without your support, I couldn't have arrived where I am today, and I will always be thankful.

Finally, to my friends who were always there when I needed support or someone to talk to, especially those in Brazil. I know the distance can be hard to maintain a close relationship, but you will always be important to my personal development and for believing in myself.

ABSTRACT

In the past decades there has been a big increase in the access to mobile phones by the population. This growth provided a useful tool to analyze mobility patterns of the population, since most of the population carries its phone to its daily routine. Understanding those patterns helps in urban planning, since it can provide insights to areas to improve public transportation, adapt the urban flux according to the time distribution of population and detect anomalies in real time. In this work, a mobile operator provided data of its users from the city of Lisbon and this data will be analyzed to detect the population behavior across the year. The results of spatial-temporal behavior of population across the year can be used to cluster the city in regions of similar patterns to define different strategies on the urban planning in those regions.

KEYWORDS

Mobile Phone Data; Clustering; Urban Mobility; Machine Learning; Unsupervised Learning.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ARIMA	Auto Regressive Integrated Moving Average
CDR	Call Detail Records
GPS	Global Positioning System
IoT	Internet of Things
IPDR	Internet Protocol Detail Records
O-D	Origin-Destination
SARIMA	Seasonal Auto Regressive Integrated Moving Average

1. INTRODUCTION

1.1. MOTIVATION

Big cities tend to have large problems when considering urban mobility. Whether in the control of the traffic flow (Alam et al., 2019), optimizing public transportation (Ma et al., 2017) or identifying anomalies in the city (Mokhtari et al., 2022), understanding the population distribution is fundamental to helping the population move. A key factor that tends to be considered in large metropolitan areas is the commuting movement.

The tendency is that job offers grow in the city center, while house renting prices are usually higher in those areas. For that reason, the urban growth tends to extend to farther places, while the job concentrations grow in the center. This situation causes a large impact on the growth of urban commuting, increasing problems in traffic (Yang, 2020), stress (Zhu and Fan, 2018), and others.

While traditionally studies relied either on surveys or traffic counters, the growth in the use of mobile phones has provided a useful resource for monitoring the population. Many studies have relied on this data to monitor urban dynamics, which will be explored in the literature review.

1.2. RESEARCH OBJECTIVES

As discussed in the motivation, this study aims to analyze the urban dynamics of the city of Lisbon. A mobile operator has provided a dataset containing anonymized mobile data collected over one year. This dataset includes information on the number of users and other relevant details within a defined grid of the city of Lisbon, without identifying individual users.

Extensive research has been conducted on existing literature regarding urban dynamics, utilizing mobile data and other tracking methods. While much of the literature focuses on specific user IDs and tracking their routes, fewer studies examine the spatial zones of the city. Additionally, there are fewer studies on clustering techniques compared to those on supervised learning models.

To address these gaps in the literature and enhance the analysis of urban patterns in Lisbon, this research has the following objectives:

1. Understand the urban patterns within Lisbon, with a particular emphasis on identifying regions with similar behavior throughout the week. Explore the spatial and temporal dimensions of the grids to discern patterns between different zones in the city.

2. Implement clustering methodologies to categorize urban zones based on shared characteristics. Provide meaningful interpretations of the defined clusters and compare different clustering strategies.
3. Interpret the results to develop strategies for urban planning. Define a scope of work that can be implemented in the city to improve the distribution of people.

1.3. THESIS STRUCTURE

This research is organized as follows:

- Chapter 2: A literature review will be conducted on past studies in urban dynamics and tracking methods, comparing different approaches.
- Chapter 3: The available data for the study will be explained.
- Chapter 4: The methodology developed for the study will be detailed to show how the results were achieved.
- Chapter 5: The results obtained in the study will be presented, with graphics and maps representing the urban structure.
- Chapter 6: A discussion of the results will focus on their implications for urban planning.
- Chapter 7: The conclusion will summarize the results and suggest possible future approaches.

2. LITERATURE REVIEW

Understanding urban spatial dynamics is crucial for effective urban infrastructure planning. One key aspect involves comprehending the flow of people, to manage traffic and help plan public transportation (Ghahramani et al., 2020). Commuting patterns, defined as the regular travel from home to work, have garnered considerable attention in studies due to the rising commuting volumes, increased time spent in traffic, and growing daily average distances in most metropolitan areas in North America and Europe (Aguilera et al., 2009). To adapt to these changing dynamics, it becomes essential to explore innovative methods for obtaining population distribution data.

Traditionally, the data has been obtained from costly and time-consuming official surveys (Yang, 2020). However, the spatial structure of modern cities has changed significantly over the last decades, largely shaped by the advancement of transportation and communication (Anas et al., 1998). Considering the communication advances in society, the access to mobile phones has become a daily resource for most people. As a result, studies are using mobile phone information to understand urban behavior.

This section will present different tracking methods to obtain location data of the population (section 2.1) and the different outcomes other related works are obtaining from this data (section 2.2).

2.1. TRACKING METHODS

To better understand urban distribution, different studies have collected data from various sources. Each approach has its advantages and disadvantages. In the following topics, the main tracking methods found in the literature will be discussed, along with some examples, to present what has already been developed in this field.

2.1.1. Survey

Commuter decisions are crucial to developing strategies for reducing and spreading the peak (Mahmassani et al., 1993). For this reason, it is important to collect data regarding the population to develop those strategies. Historically, the most traditional method for acquiring individual-level data on urban distribution is through surveys, commonly referred to as primary or small data (Yu et al., 2020).

Before the growth in the use of mobile phones and, eventually, smartphones, the data collection about the population was done manually through interviews, either in person or by calling residential phones. This type of data was useful for the time, as there was no more automated way of collecting information from the population. Even now, many researchers rely on surveys to collect information from users, either to complement data or as the only source.

One notable advantage of surveys is the flexibility they offer in tailoring questions to specific information needs and getting more precise data. However, to get a good representation of the population, the higher the number of people interviewed, the better, and getting this amount of people can be time consuming and costly.

Periodically, the government also releases census information about the population that can be used for research. Even though census data can represent all the people of a community, its release is not yearly, and the data may not be up to date for some researchers, especially when society is changing fast. Also, not every question is answered on a census, so the researchers might need to implement a survey to complement the data.

2.1.2. Call Detail Records (CDR)

The advent of big data has revolutionized the study of large-scale human behavior, particularly in the domain of human mobility (Kung et al., 2014). Mobile phones are one of the fastest growing technologies in the world and information generated by them already provides insights into official statistics (Jahani et al., 2017). In that field, call detail records or CDR are one of the methods that can be mostly found in the literature.

Each time a mobile user makes or receives a call or text message, the nearest cell tower registers the cell tower location, the caller and receiver ID, timestamp, and other information (Phithakkitnukoon et al., 2017). The service providers collect this data for operational, planning and billing purposes (Becker et al., 2011) and are required for legal compliance, thus it is stored for an extended period (Ghahramani et al., 2020).

Since this data is already stored by mobile operators, it is used to deduce models for large geographic areas with diverse populations (Zhang, 2014). If the population in the study is distributed evenly between the different mobile operators by gender, age and social background, this subset can be a representation of the region for a study.

One of the key issues in using CDR is that it requires the user to make or receive a phone call or text message. This causes part of the data to be sparse between users. Another concern about this method is related to the privacy of the person. Most of the data used in studies are anonymized. However, it can still be used to track the most frequent places that people make phone calls. Lastly, the uneven distribution of cell towers in a city can provide some biased information, especially in more rural areas.

2.1.3. Global Positioning System (GPS)

The Global Positioning System (GPS) is a geolocation service relying on satellites, allowing a receiver (such as a smartphone) to determine its location based on signals received from satellites within its field of operation. GPS records provide valuable information, including latitude, longitude, altitude, and timestamp for an individual's position.

Modern smartphones already have built-in GPS devices. Like the CDR approach, GPS can provide relevant user information throughout the daily journey. However, it may encounter limitations in underground environments and face interference in regions with high building density (Qiao et al., 2017). Moreover, GPS provides specific coordinates, elevating privacy concerns.

Some applications found in the literature include, for instance, a study aimed at detecting meaningful places in daily life that can utilize historical GPS locations to output a geometry of a set of places (Zhou et al., 2007). A proof of concept can also be developed for a traffic monitoring system based on GPS, detecting users' velocity at different points on the freeway (Herrera et al., 2010).

One thing to consider about GPS over CDR is that GPS can be applied to other devices that are not mobile phones. The Internet of Things, or IoT, is the intercommunication between devices and the internet. Although it is a broad concept, it is related to communicating equipment from the quotidian with sensors and providing this information to a monitoring system or other applications. Since GPS is already being used in vehicles, such as cars, buses and taxis, and other devices, like smart watches, there are studies about urban distribution that don’t take into consideration the use of a smartphone.

Smart cities are the goal for the future of citizen habitation. The location provided by GPS in cars can develop a predictive vehicle ride sharing recommendation to change the way citizens commute (Anagnostopoulos, 2021). Wearable devices have already been used in some cities to monitor patients during the pandemic of COVID-19 (Al Bassam et al., 2021). Watch-like devices were also used to monitor users' daily routines to track patterns both individually and collectively (Neuhaus, 2010).

2.1.4. Internet Protocol Detail Records (IPDR)

While previous approaches provide a good understanding of the urban distribution of users in a city, there are still some difficulties in collecting this data. CDR can provide a good estimate of user location, but it depends on the user making or receiving a call or text message. GPS does not require the user to make a call, but its data can have some interference in some areas and GPS applications consume devices energy quickly.

To improve this situation, recent studies are collecting information from the users accessing the internet through data networks (2G/3G/4G/5G). Collecting this data while citizens are moving through the city has many advantages (Qiao et al., 2017). Since people are using data networks more frequently than voice services and use data for gaming, music, news, chatting and others (Zhang, 2014).

While the literature does not use a common name for this approach, IPDR will be used in this study for comparison with CDR. Every time a user accesses the internet, the Internet Service Provider (ISP) collects some data for billing and accounting purposes (Lawgic, 2022).

So, whenever a citizen accesses the internet in their daily journey, the ISP has information about the cell tower the device is connected to and its location.

Similar privacy concerns, as with the other approaches, must be considered for this study. Other problems with this approach are that many devices change from mobile data to Wi-Fi, which can provide some hidden locations. It is important to note that users tend to connect to Wi-Fi in more important places, such as their workplace or their house.

For this study, mobile data information was collected from both CDR and IPDR datasets. Since the growth in internet usage is faster than voice, it is important to analyze this data in the study. However, voice data can be particularly important in workplaces that rely on calling customers or suppliers or even for some population that prefer calling to accessing the internet.

2.1.5. Others

While the majority of studies in this research employ the approaches described above, there are some less commonly used methods to track urban distribution.

When studying the public transportation system, smart cards are being used by the population. Whenever a citizen accesses a bus or enters or exits a subway station, the smart card is used, and the information is transmitted in reference to the place. This can be used to identify commuting patterns via public transportation (Ma et al., 2017).

Before the advent of mobile internet, the traffic monitoring architecture mainly consisted of loop detectors, cameras, and radars (Herrera et al., 2010). Video cameras placed at intersections could be used to measure traffic and turning movements for each direction in the period recorded (Kyte et al., 1993). Sensors placed on city roads could measure traffic to develop a regression model (Alam et al., 2019). Traffic counters can also be used together with mobile data to develop intelligent road traffic statuses (Demissie et al., 2013).

2.2. MAIN OBJECTIVES

Once the data is obtained, many different pieces of information can be retrieved. The literature shows a diverse range of objectives when studying urban areas. In this section, there will be a brief description of some methods used in the literature. Since the main goal is not to explain the methods in detail but to understand what researchers are obtaining as results, this section will primarily describe these methods.

2.2.1. Origin-Destination Matrix (O-D Matrix)

Commuting patterns are usually represented by O-D Matrices (Frias-Martinez et al., 2012). As the name suggests, these matrices represent geographical information on the population flow from an origin to a destination. When analyzing commuting patterns, the important thing is to understand the journeys from home to work and vice versa.

When building these matrices, we need to identify the regions considered as residential and workplaces. When a study is based on a survey, the question can be directly asked. However, when using data derived from mobile phones or similar devices, there must be other strategies. Usually, work hours are from 9:00 a.m. to 5:00 p.m., with a small variation according to the region, city, or job. So, when a user ID has multiple occurrences on weekdays in the same area during this period, this region is treated as a workplace. Similarly, most people sleep at home from 11:00 p.m. to 6:00 a.m., so activities that tend to repeat around those times multiple times on the week are considered residential.

A model using O-D matrices can be developed to analyze changes in commuting patterns based on counter scenarios (Yang, 2020). This model can compare the actual changes in the commuting patterns with changes that increase the number of houses in central areas and jobs in not central areas (decentralization of jobs). These are possible changes discussed to attempt to mitigate the commute time.

One of the main concerns when studying this strategy is that this data can represent sensitive information, as it tracks a user ID in their residential and workplace area. Also, this strategy cannot cover the movements inside the same area represented in a matrix (Ghahramani et al., 2020).

2.2.2. Classification Model

One of the categories that Machine Learning can be divided into is Supervised Learning. This division represents the predictive models, where the user inputs data and the model will produce an output. When this output is a defined value or category, those are called classification models.

While there are many different classification models, the goal of this study is to analyze the outcomes obtained from the literature. A simple classification model can be developed to determine if the users are commuters or not (Yu et al., 2020). The complexity can be increased to determine other variables. For instance, a model can be developed to predict the traffic levels in specific points of the city (Demissie et al., 2013).

Models can be associated with other strategies. A user profile can be developed by analyzing past data and, with that, creating a classification model to determine future locations (Anagnostopoulos, 2020). Even information related to city demographics and statistics can be developed, to determine user characteristics, such as gender and age based on the information obtained from mobile phones (Jahani et al., 2017).

The main issue when considering using classification models is that the data must be labeled. So, if the goal is to retrieve unknown information or group data, other models should be considered. For this study, the data is unlabeled, and the classification models from the literature mostly served as inspiration for developing our model.

2.2.3. Regression Model

Still in the group of Supervised Learning, when the output from the model is one value from a continuous range of values, this is called a regression model. While a similar concept to the classification models, the approaches tend to be slightly different.

One simple regression model can be used to estimate the distance or time spent commuting by a user (Kung et al., 2014). It can also be used to predict the total volume of traffic in specific points of the city (Alam et al., 2019).

Other approaches tend to consider individuals' lifestyles. For example, a regression model can be associated with personal characteristics of the user to calculate the probability of the individual commuting to peripherical zones (Zhao et al., 2011). The level of happiness of an individual can also be measured based on the daily choices of commuting by the individual (Zhu and Fan, 2018).

Although regression models can be quite useful, there are some constraints. One thing to consider is that the models only represent part of the real-world data, since not every variable can be used in the same model. Outliers also tend to impact the results, which may cause inaccurate conclusions.

2.2.4. Clustering

In contrast to Supervised Learning, Unsupervised Learning does not require data to be labeled or to have a pre-determined output. The algorithms identify hidden patterns in data to label inputs together. When the input data is grouped into similar categories based on their characteristics, this is called clustering.

There are different methods of clustering. One of the most used methods is k-means, where the data is grouped into k groups, defined by the data distribution in space. Apart from the strategy used for clustering, the approach can change from clustering users, regions, and others.

Different users' patterns can be determined by clustering similar cellphone usage on the days of the week (Becker et al., 2011). By defining the daily routine, the users can also be clustered by their daily routine (Jiang et al., 2012).

On the other hand, regions can be grouped into clusters based on the number of users in the area. By monitoring the city and the usage in certain areas across the week, collective human activity can be determined (Sagl et al., 2014). By determining users' job activity points and users' residences, the regions in a city can be grouped according to the ratio between the values (Yu et al., 2020).

Considering the dataset available, which is divided into location, features, and intervals, a similar study (Chidean et al. 2023) was found that clustered the regions in Milan. This study

presented a methodology where a distance metric was defined over the features and temporal range available, allowing clustering techniques to be applied to divide the city.

When using clustering techniques, one of the issues is that, since the data is not labeled, the significance of the groups must be identified by the researcher. Since there are many different clustering techniques, it is hard to identify the best method to choose and the number of clusters to select.

The data available for this study is represented in a grid format over Lisbon. Since there is no information based on a particular individual, clustering techniques are useful to identify regional similarities. For that reason, the study will focus on clustering regions to better understand the commuting patterns over the time interval in the city.

2.2.5. Time-Series

Time series is a sequence of data points that occur in sequence over a period of time. Since most of the mobile phone tracking methods register the timestamp of the occurrence, a time series analysis can provide insights into the dynamics of user behavior over distinct time frames. It enables the identification of recurring patterns, trends, and anomalies, contributing to understanding how mobile phone usage and commuting activities vary over time.

By dissecting these temporal patterns, researchers can unravel commuting behaviors and uncover peak commuting times, potentially informing urban planning strategies. Additionally, time series analysis facilitates the identification of seasonality and long-term trends in commuting patterns.

While most studies have access to the temporal information of users, most of them focus on the spatial analysis, using the temporal information as a complement (for example, determine home and workplace by the time the user is active).

The time series analysis can be associated with other models. For instance, a regression model of traffic can be associated with the time to detect the flow of people (Alam et al., 2019). It can also be used to compare different clusters of users over time to try to build meaningful clusters (Zhang, 2014). It can also be used to detect anomalous behavior of users in the city (Mokhtari et al., 2022).

2.3. SUMMARY REVIEW

As discussed in this section, there are numerous tracking methods and objectives when analyzing urban population displacement. Table 2.1 summarizes the main studies discussed in this research, focusing on similar topics. Even though the range of topics is broad, the main themes tend to revolve around the same fields. To extend the research on Clustering Methods and deepen the study on IPDR data, this study decided to cluster Lisbon zones based on mobile data.

Table 2.1 – Summary of Literature Review

Article	Tracking Method	Main Objective
Alam et al., 2019	Other	Regression
Anagnostopoulos, 2021	GPS	Classification
Becker et al., 2011	CDR	Clustering
Chidean et al., 2023	CDR; IPDR	Clustering
Demissie et al., 2013	CDR; Other	Classification
Frias-Martinez et al., 2012	CDR	O-D Matrix
Herrera et al., 2010	GPS; Other	Regression
Jahani et al., 2017	Survey; CDR	Classification
Jiang et al., 2007	Survey	Clustering
Karikoski and Soikkeli, 2013	CDR; Other	Classification
Kung et al., 2014	CDR; GPS	Regression
Kyte et al., 1993	Other	Classification; Regression
Ma et al., 2017	Other	Clustering
Mahmassani et al., 1993	Survey	Other
Martin et al., 2018	Survey (Census)	Classification
Mokhtari et al., 2022	CDR	Classification; Clustering; Time Series
Neuhaus, 2010	GPS	Other
Phithakkitnukoon et al., 2017	CDR	O-D Matrix; Classification; Regression
Qiao et al., 2017	IPDR	Classification; Clustering; Time Series
Sagl et al., 2014	CDR; IPDR	Clustering
Thuillier et al., 2018	CDR	Clustering
Xu et al., 2016	CDR; IPDR	Time Series
Yang, 2020	CDR; IPDR	O-D Matrix
Yu et al., 2019	CDR	Classification; Clustering
Zhang, 2014	CDR; IPDR	Clustering; Time Series
Zhao et al., 2011	Survey	Regression
Zhou et al., 2007	GPS	Clustering
Zhu and Fan, 2018	Survey	Classification; Regression

3. STUDY DATASET

The dataset for this study was provided by a major mobile operator in Lisbon, under a non-disclosure agreement, regarding the available data. It comprised two main components: a shapefile containing geographical information about the study area, as detailed in Section 3.1; and multiple CSV files containing mobile data information, detailed in Section 3.2.

3.1. LISBON GRID INFORMATION

The scope of this study was limited to the city of Lisbon. The city was subdivided into grids, as represented in Figure 3.1, based on the shapefile provided. The dataset for the geographic information of Lisbon has the features described in Table 3.1 below:

Table 3.1 - Description of shapefiles features

Feature	Description
WKT	representation of the grid coordinates in well-known text format
Geometry	representation of the grid coordinates
Latitude	grid centroid latitude
Longitude	grid centroid longitude
Grelha_y	y-axis alternative coordinate
Grelha_x	x-axis alternative coordinate
Grelha_Per	grid perimeter
Grelha_Are	grid area
Freguesia	designation of the current parish of the centroid of the square
Freguesias	designations of the former parishes of the centroid of the square
Nome	common designation of the area characterized by the grid
Dicofre	unique identifier of the parish
ID	unique identifier of the grid.

Upon analysis of the dataset, some insights emerge. Firstly, all grids exhibit uniform characteristics, with an identical area of 40000 m² and perimeter of 800 m. This uniformity results from the city's division into grids of the same shape and size, each one a square with sides measuring 200 m.

Overall, the city was divided into 3743 grids, representing 24 parishes called “Freguesia”. The latitudes vary from 38.692094° to 38.797189°, over 76 different rows, while the longitudes vary from -9.230417° to -9.090280°, over 79 different columns.

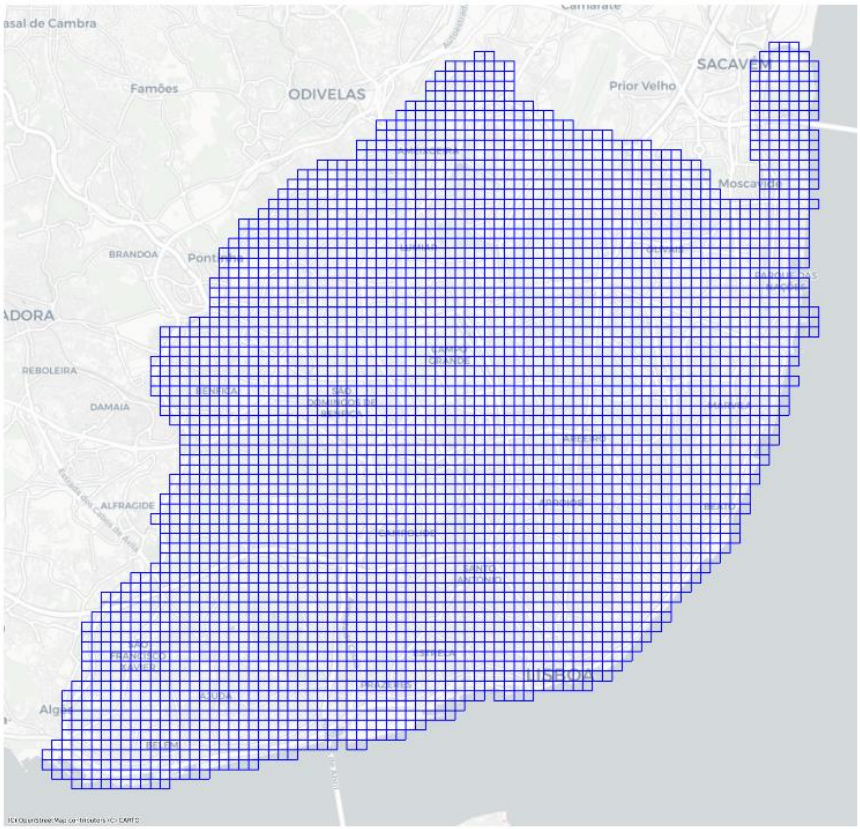

Figure 3.1 – Lisbon grid division as available on the dataset.

3.2. MOBILE INDICATORS

The mobile operator provided multiple CSV files covering the period from September 2021 to August 2022, which combined provided the dataset. The information provided contained records every 5 minutes during the period described. The dataset structure resembles a cube of features, as depicted in Figure 3.2, where one dimension represents space (the grid the record is related to), another represents time (the timestamp period of the record), and the third dimension represents the recorded values of the features.

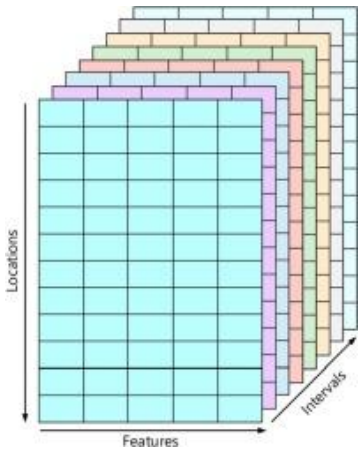

Figure 3.2 – Dataset cube representation. (Chidean et al., 2023)

The dataset contains numerous features, which are detailed in Table 3.2 below:

Table 3.2 – Description of mobile indicator features

Feature	Description
Grid_ID	represents the ID of the grid, that correlates with the shapefile
Datetime	timestamp of the record containing the date and time
C1	number of distinct terminals in the grid
C2	number of distinct terminals, roaming, in the grid
C3	number of distinct terminals that remained in the grid
C4	number of distinct terminals that remained in the grid, roaming
C5	number of different terminal inputs in the grid
C6	number of outputs of distinct terminals in the grid
C7	number of different terminal entries, roaming, in the grid
C8	number of different terminal outputs, roaming, in the grid
C9	number of distinct terminals with active data connection, in the grid
C10	number of separate terminals with active data connection, roaming, in the grid
C11	number of voice calls from the grid
D1	top 10 countries of origin of roaming terminal devices
E1	number of voice calls that ended in the grid
E2	average downstream rhythm of the grid, in the grid
E3	average grid upstream rhythm
E4	quad downstream peak rhythm
E5	quadruple peak rhythm
E6	top 10 Applications
E7	duration of minimum stay within the grid
E8	duration of the average stay in the grid
E9	duration of maximum stay within the grid
E10	number of devices that share the connection in the grid

About the features, the goal of the study is to perform an analysis of the spatial distribution of users across the city of Lisbon. Therefore, non-numerical features (D1 and E6), as well as the features regarding the rhythm of data (E2, E3, E4, and E5), were not used.

Upon analysis of the dataset, the first notable observation was the difference in file sizes for the months. The months of February and March were relatively smaller than the others. During data exploration, it was discovered that there was a gap in the data between February 8th and March 18th. This issue will be discussed in the Methodology section.

Continuing the data exploration, Table 3.3 presents the statistical description of the features, including mean, standard deviation, minimum, median, and maximum values.

Table 3.3 – Average description of the features

Feature	Mean	Std	Min	50%	Max
C1	158.736519	227.747061	0	84.91	39163.92
C2	7.756876	22.206896	0	1.25	2998.21
C3	123.930392	188.221639	0	63.23	37642.48
C4	6.102592	18.001642	0	0.73	2889.72
C5	67.276377	109.254882	0	32.67	17114.85
C6	76.618418	108.526655	0	43.03	11533.88
C7	3.157411	9.918534	0	0	517.49
C8	3.258743	9.225407	0	0.44	473.83
C9	154.320187	222.227093	0	82.33	39020.47
C10	7.676547	21.957170	0	1.25	2987.60
C11	3.553527	11.830074	0	0	6913.24
E1	1.057323	6.159358	0	0	3744.54
E7	0.342548	4.150723	0	0	300
E8	6.101987	9.193805	0	5.06	300
E9	62.796418	77.149309	0	34.78	300
E10	0.000878	0.085761	0	0	29.47

Among the features starting with C, C1 stands out as the largest set, encompassing all other C features as subsets. Consequently, it exhibits the highest values across all features and serves as the primary reference for the study. C2 is also notable for being the largest set concerning users utilizing data on roaming.

Key insights from the table reveal that features C11, E1, E7, and E10 have low averages, with over 50% of their values being null. Therefore, E1, E7, and E10 were excluded from the study. C11, however, was retained due to its slightly higher average and its equivalence to CDR (as discussed in the literature review). CDR has been pivotal in numerous studies on mobile data distribution in cities, although the number of calls has significantly declined with the widespread adoption of 3G, 4G, and 5G networks.

A useful tool for data exploration is the correlation matrix. In Figure 3.3, the average correlation matrix of the features across the months can be seen. As shown, C1 has a high correlation with C3, C5, C6, and C9, while C2 has a high correlation with C4, C7, C8, and C10. When using features that are highly correlated with each other, the analysis tends to be biased towards those features, as if they were counted multiple times. To simplify the study

and avoid this bias, it was decided to keep only C1 and C2 as features, since the features highly correlated with them are subsets of C1 and C2.

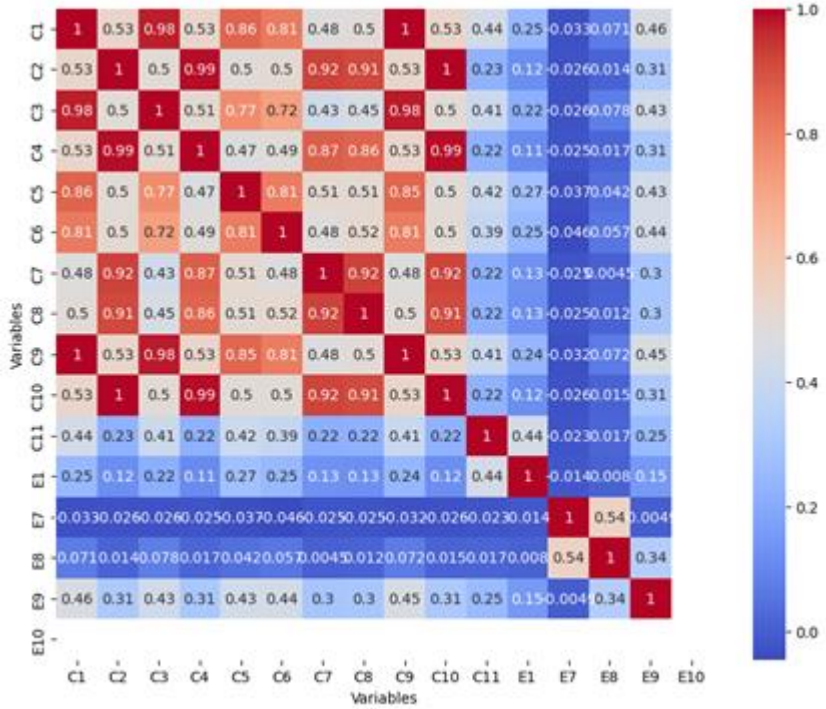

Figure 3.3 - Correlation matrix of features.

The final objective in the data exploration was to identify possible outliers in the data. Figure 3.4 shows a boxplot of the features presented in the dataset for a given month. As it illustrates, a significant portion of the data falls outside the standard boxplot range (interquartile range extended by 1.5 times the interquartile range above and below the quartiles).

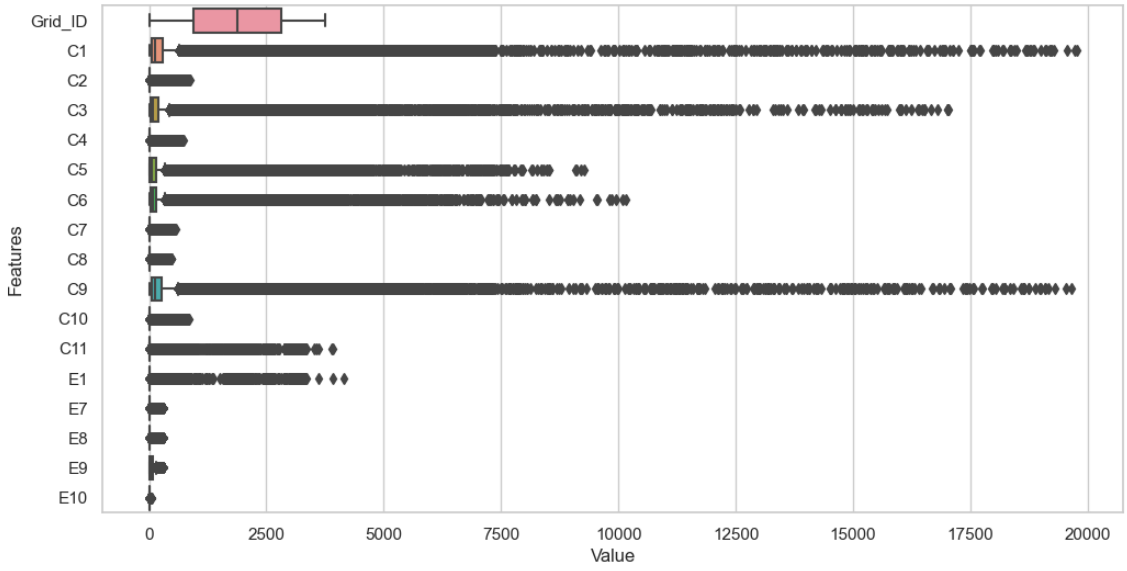

Figure 3.4 – Boxplot of features for a given month.

Figure 3.5 displays a time series plot for the C1 feature in a random grid for the first week of the dataset. As shown in the figure, several outliers fall outside the range between the mean and 1.5 times the standard deviation. Ideally, the interval to be considered should be manually analyzed to determine the best range for removing outliers. However, given the large size of the dataset and the varying behavior of different grids, a comprehensive manual analysis is impractical. The strategy used in this study to mitigate the impact of outliers will be discussed in the Methodology section.

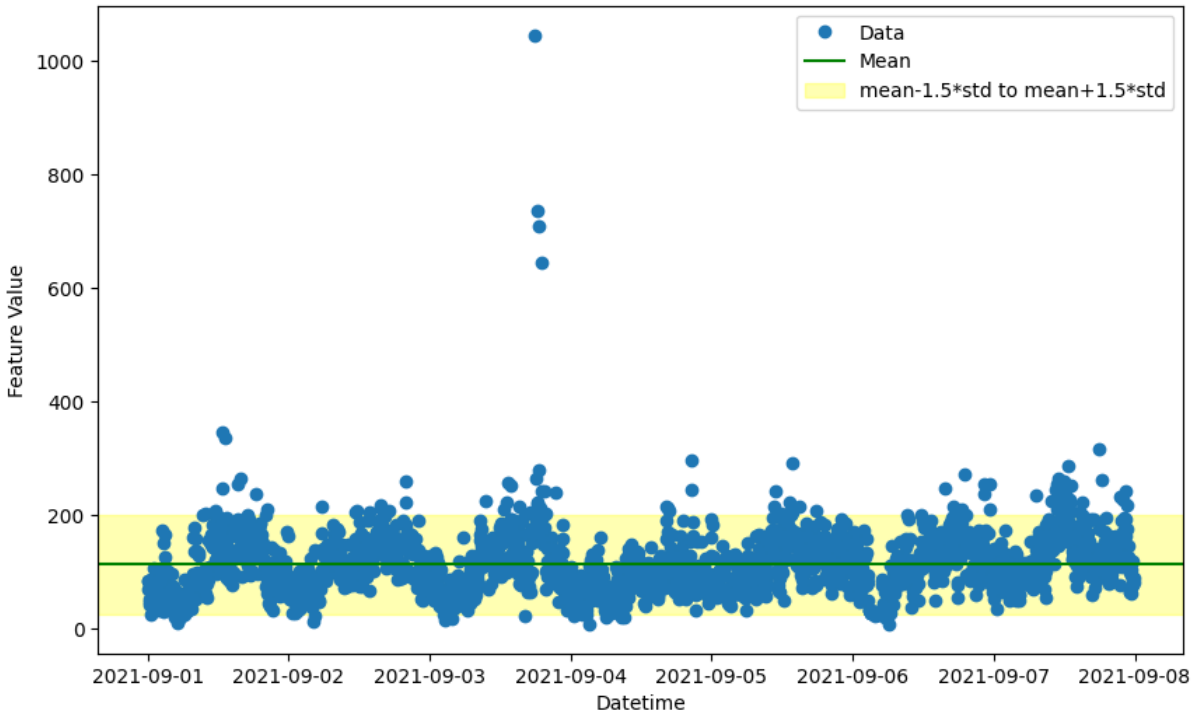

Figure 3.5 – Time Series of C1 for a random grid in the first week of dataset.

4. METHODOLOGY

Following the concepts discussed in section 2 and the data available described in section 3, this research aims to compare the behavior of mobile phone users in the different zones of Lisbon, clustering the zones based on predetermined characteristics. The study was conducted using Python and typical machine learning libraries

The CRoss-Industry Standard Process for Data Mining, better known as CRISP-DM, is the standard methodology to guide the most common steps in data mining projects (Martínez-Plumed et al., 2021), and it is illustrated in Figure 4.1.

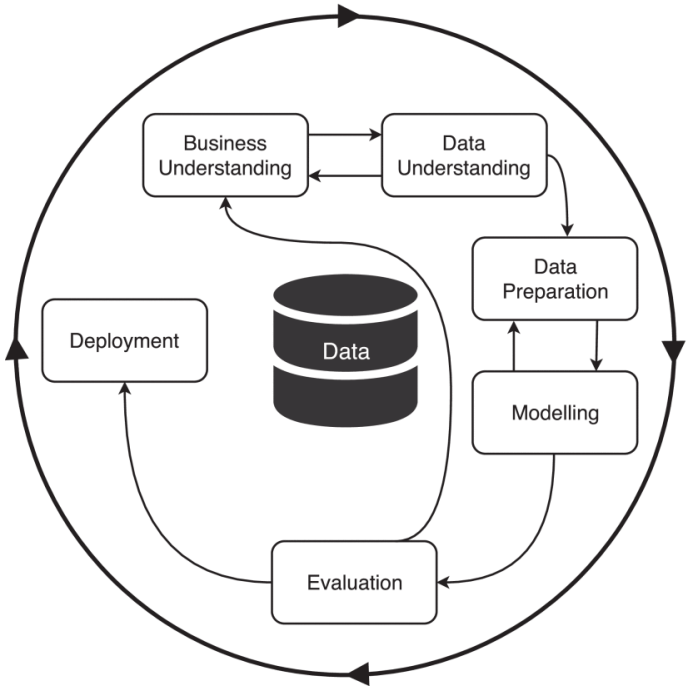

Figure 4.1 – CRISP-DM methodology representation (Martínez-Plumed et al., 2021).

4.1. BUSINESS UNDERSTANDING

Numerous studies have focused on population distribution, commuting patterns, and other factors to develop strategies for urban planning. With the advancement of mobile phones in the daily routine of the population, these devices serve as constant tracking tools that aid in monitoring population movements and behaviors. This phase of the Methodology was conducted in the Literature Review, and more details about the business understanding can be found in that section.

The goal of this study is to cluster the regions of Lisbon based on population patterns observed across different days. Once the clusters are obtained, it is crucial to understand and evaluate how well the zones were divided and what each cluster represents. This understanding will help in developing strategies for effective urban planning.

4.2. DATA UNDERSTANDING

The datasets for the study were presented in Section 3. The first dataset is related to the region of Lisbon and its division into grids, while the second dataset contains information related to mobile phone data within these grids. In this step of the study, it is important to conduct an exploratory analysis of the data to better understand how to proceed. This part of the process involves obtaining insights to make decisions about the study, such as strategies to be employed in data preparation.

4.3. DATA PREPARATION

Data preparation is a vital step in most machine learning studies, with the goal of improving data quality and eliminating irrelevant information from real-world datasets. The available dataset comprised multiple CSV files, totaling over 60 GB of data.

The first requirement was to reduce the number of features available in the dataset, as multiple features were not within the original scope of work. Additionally, some features represented a subset with a high correlation to the original set. Therefore, those features represented redundant data and were removed.

After the removal of some features, the goal was to fill missing values. When comparing multiple time series, ideally, they must be of the same length to represent the same interval. Although there are strategies to deal with time series of different lengths, it requires fewer computational resources and is more intuitive to fill the missing values.

Another crucial point to address in the dataset is the presence of outliers. Clustering techniques that require the calculation of distance between points, such as K-Means, are highly sensitive to outliers. Identifying and appropriately handling outliers is essential to ensure a good result.

Finally, there are some correlations between the available features that can provide new values and generate new features. This process is called feature engineering, where new features are generated from the ones already available. In the next topics, each process of data preparation will be detailed.

4.3.1. Filtering Features

To reduce the data, features that were not within the scope of work were removed. The study did not intend to analyze textual data, such as the countries of origin from roaming devices or the most used mobile apps. Additionally, the rhythm of the data was not analyzed since the primary goal was to understand the urban flux of people, not the quality of the data signal.

Features that consisted mostly of zeros (over 75% of the values) were removed. The remaining data contained multiple features that were correlated with each other, as shown

in Figure 3.3. To reduce redundancy and focus on key points of the analysis, some features were selected. The features chosen for further analysis were C1, C2, C11, E8, and E9.

4.3.2. Filling Missing Values

It is important for the clustering model that every grid has the same amount of data points. To fill missing data in a time series model, there are several different techniques available.

One common strategy to fill missing data is interpolation. This strategy tries to fit a function between two known points to fill the missing gap between them. This strategy is useful when there is a small interval between points (for instance, the value for 10:15 can be the average between 10:10 and 10:20, considering a constant rate for increase or decrease). However, as the interval grows, more complex functions are required to understand how the missing points will behave, even cyclic behaviors.

Other strategies may apply time series analysis. Creating a model to identify the time series behavior, such as a SARIMA (Seasonal Autoregressive Integrated Moving Average) model, could help to fill the missing data. However, these models require high computational resources to apply.

To simplify the study, the strategy adopted was to divide the week into equal intervals from Monday to Sunday. As population behavior across the week tends to repeat, the value could be filled by the values over the same equivalent period in other weeks. This strategy is similar to an imputer commonly used in machine learning projects, and the imputation strategy adopted was the 'median', as a way to avoid outliers influencing the result.

4.3.3. Removing Outliers

It is very important to have a strategy to deal with outliers in a project. Outliers are points in the dataset that deviate from the general trend of the data. This could happen either due to a mistake in reading the data or because of an actual anomaly in the data. For instance, an error could occur when the cellphone towers detect the wrong number of devices in a grid. Another situation could be a traffic congestion, which would represent a real anomaly, increasing the number of devices in the area.

To deal with outliers, the researcher should decide whether to keep, remove, or change the value of the outlier. No strategy is perfect, and the researcher should choose the one that offers the most benefits.

One common strategy for detecting outliers is to use the interquartile range (IQR), which is the distance between the 25th percentile (Q1) and the 75th percentile (Q3) of the values for a given feature in the dataset. To identify outliers, lower and upper bounds are established by subtracting 1.5 times the IQR from Q1 and adding 1.5 times the IQR to Q3, respectively. Using this strategy, many data points were not within the interval. It is not recommended to remove them if they represent a significant portion of the total data.

When dealing with time series data, another strategy is to decompose the time series into trend, seasonal, and residual components. The general behavior of the time series should be a combination of the trend and seasonal components. For the residuals, they can either be removed to "clear" the data of outliers, or a threshold can be established for the maximum acceptable residual for the study.

However, the study faced a significant challenge when comparing multiple grids. A value could be considered an outlier in one grid but represent expected behavior in another. Additionally, the dataset was very large, containing data every 5 minutes across one year, which required extensive processing.

To both reduce the amount of data to process and mitigate the outliers, the dataset was aggregated to the median by hour. The goal of grouping the data by hour was to reduce the dataset to one-twelfth of its original size, as each hour contains twelve “five-minute intervals.” Using the median to group the data by hour helps reduce the influence of outliers. A single data point could deviate from its typical behavior, but the median selects the value around the middle, thus not counting the outlier. If another metric, such as the average, were used, the outlier would influence the result.

4.3.4. Feature Engineering

In machine learning projects, it is often necessary to generate new features based on the available data or to modify existing ones. This process is known as feature engineering. From the original dataset, the features available after filtering were C1, C2, C11, E8, and E9. For this study, two datasets were generated to perform two different clustering techniques, which will be detailed in Section 4.4. The process of modifying the features to generate each dataset is explained below.

The first dataset is the “Time Series Dataset.” The goal of this dataset is to analyze how features behave over time. One issue with this approach is that two features shouldn’t be compared together over time. For that reason, the study focused on C1 for the analysis, as it is the largest set available and more generically represents the population distribution.

To calculate this dataset, only the C1 value was maintained. However, when performing operations over time series, the complexity of the operation increases with the length of the time series. Currently, the available data consists of one year of data, divided hour by hour. To simplify the analysis while preserving weekly patterns, the week was divided hour by hour, starting from Monday at 0:00 (point 0) to Sunday at 23:00 (point 167), totaling 168 points (24 hours per day over 7 days in a week).

Once the new data points were defined, the data was aggregated by the same time periods. Thus, the C1 values were averaged over the same time periods across the year, resulting in an average value for each day and hour of the week.

After that, the values were normalized over the week for every grid. This means that, in each grid, the lowest value of the week was set to 0 and the highest value of the week was set to 1. This normalization was done to distribute the weekly values for each grid within the same interval, ensuring that, when performing clustering techniques, the proportion of the distribution across the week in each grid is taken more into account over the differences in population distribution between different grids.

The second dataset is the "Features Dataset." While the first focused on the behavior of features over time, this dataset focuses on the feature values across different grids. This strategy provides greater flexibility in choosing various features to compare without considering the time component.

Given this flexibility, some features were generated to try to identify new patterns not available in the original features. The first feature generated was the ratio of C2 to C1, representing the proportion of terminals in the grid that are roaming. This can be useful for understanding the regions in Lisbon that are most touristic, as there is a higher proportion of roaming devices.

Another feature generated was the C1 differential, representing the change in C1 between consecutive timestamps (with the first record set to 0 by default). This feature helps detect regions with larger fluctuations in population between consecutive hours.

Additionally, a feature called "day peak" was created. This feature calculates the hour when the number of terminals in the grid (C1) is highest for each day. The goal of this feature is to differentiate grids that tend to have peaks earlier in the morning from those that peak later in the evening.

Finally, two categorical features were defined. One feature separates weekdays from weekends, allowing for a comparison of behaviors between them. The other feature identifies the "worktime", defined as the hours between 10:00 and 18:00.

Once all the features were generated, the Features Dataset was calculated. For this, the numerical features were normalized to ensure that differences in scales between features do not influence the results. The following metrics were then calculated over the entire time period: the average values of the original features (C1, C2, C11, E8, and E9); the average day peak; the maximum and minimum C1 differentials (the maximum increase and maximum decrease); the average C2 to C1 ratio; and the average values of C1 and C2 for weekdays, weekends, and worktime.

4.4. MODELLING

Once the data preparation is complete, the next step is to model our data. For this study, the chosen model involves applying clustering techniques to the Lisbon grids. As explained in Section 2.2.4, clustering will group the regions based on similar behavior. In the previous

section, it was explained that two datasets were generated. This was done to apply two different clustering strategies, which will be detailed in the following sections.

4.4.1. Clustering by Time-Series

The first method for clustering used in the study is a temporal proximity-based method. This method aims to identify similar time series in shape, as shown in Figure 4.2. However, this method requires more processing since it will measure the distance point by point in the time series. For that reason, the time series was averaged by the same day of week and hour, so we limit our time series for just one week, as an average of the other weeks.

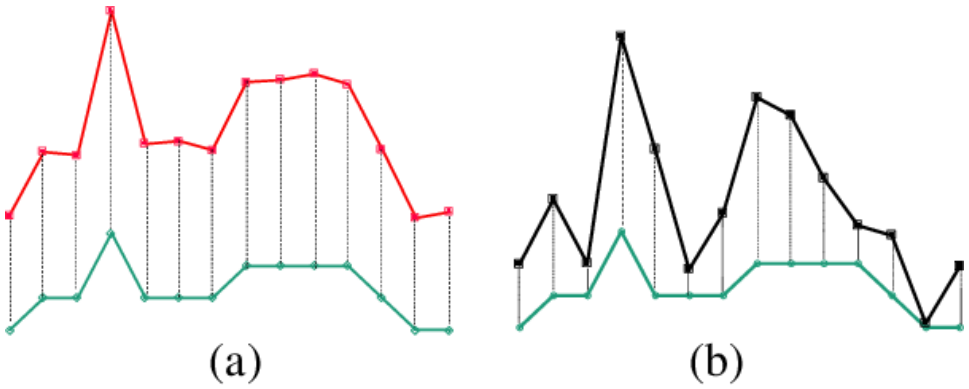

Figure 4.2 - Illustration of the distance between two time series (Puri et al., 2022).

Despite the dataset representing a single feature across different timestamps, traditional clustering methods are applicable. For this study, two clustering techniques were utilized: k-means and hierarchical clustering.

K-means is an algorithm that generates k random centroids and measures which data points are closer to each centroid. Then, the centroids are updated according to the points defined closer to each centroid. The algorithm is iterated until a stopping condition is applied, such as a change in the centroid smaller than a threshold or a certain number of executions of the algorithm. Some disadvantages of this method are that the number “k” of clusters must be defined before applying the technique. It is also highly sensitive to outliers and its result may vary depending on the starting conditions (randomly generated).

Hierarchical clustering can be either agglomerative or divisive. For this study, agglomerative clustering was applied. Initially, each data point starts as a single cluster. Points are then merged based on their proximity to each other, iteratively combining the closest clusters until all points belong to a single cluster. Once all clusters are agglomerated into one, a dendrogram is typically generated to visualize the distances used in merging the clusters. A threshold is then defined from the dendrogram to separate the clusters.

Once performed each clustering technique, the results should be interpreted and compared between each other, to identify advantages and disadvantages from each strategy.

4.4.2. Clustering by Features

The second method for clustering the Lisbon zones was based on the features dataset generated. The process to generate this dataset was explained on topic 4.3.4. The idea for this type of clustering is illustrated in figure 4.3, where it shows that the raw time series had its features extracted into a new dataset, that reduces its dimensions to perform typical clustering algorithms.

Figure 4.3 – Feature Extraction. Retrieved from: https://heka-ai.medium.com/time-series-clustering-b84bcaaa63ac

Since a lot of features were generated, a greedy algorithm could be applied to determine the best features on the cluster process (fmarthoz, 2021). However, it took a significant amount of time and computational resources to apply the feature selection. To simplify the study, it was decided to retain all the original features.

After that, the next step is to perform clustering techniques. To maintain consistency with the methods used in time series clustering, both k-means and hierarchical clustering were applied. Once the clusters are defined, an interpretation of the clusters will be conducted to understand the behavior of users within each cluster. The results should be compared between both methods (time series and features) to identify advantages and disadvantages.

4.5. EVALUATION

The next step in the CRISP-DM methodology is to evaluate the results. A common metric used to evaluate the performance of clustering techniques is the silhouette score, which measures how similar a data point is to its own cluster compared to other clusters. A high silhouette score indicates that the data points are well-separated into distinct clusters.

Once the clustering has been evaluated, the subsequent step is to apply the researcher's business understanding to gain deeper insights into the clusters and develop strategies for urban planning in the city of Lisbon. These strategies could enhance public transportation, stimulate growth in underdeveloped commercial zones, and potentially reduce commuting times for the population.

Finally, while the model deployment is essential for city planners, it falls outside the scope of this study.

5. RESULTS

In this chapter, it will be discussed the results obtained from mobile data analysis in the city of Lisbon.

5.1. CLUSTERING BY TIME SERIES

As detailed in the Methodology section, one of the techniques employed for data analysis was clustering by time series. Clustering time series data offers the advantage of preserving temporal information within the dataset, although it presents challenges such as handling data from a single feature and sensitivity to missing data. Given the dataset's characteristics, it was decided to cluster the attributes using the C1 feature, which represents the largest set among the original features.

Before performing the clustering techniques, it is important to understand the data preparation, discussed in topic 4.3, and how the time series dataset was generated, as explained in topic 4.3.4.

5.1.1. K-means

Once the data is correctly grouped, the cluster analysis can begin. The first method used was K-means. Since this method can be impacted by different scales in the time series, each one was normalized (scaled between 0 and 1) to ensure uniformity within the weekly intervals.

One parameter that needs to be defined in the K-means algorithm is the number K of clusters. The silhouette score is a useful metric that measures how closely elements within a cluster resemble each other compared to elements in other clusters. Figure 5.1 shows the silhouette scores applied to the dataset. From this, it is evident that 2 clusters perform significantly better than other numbers.

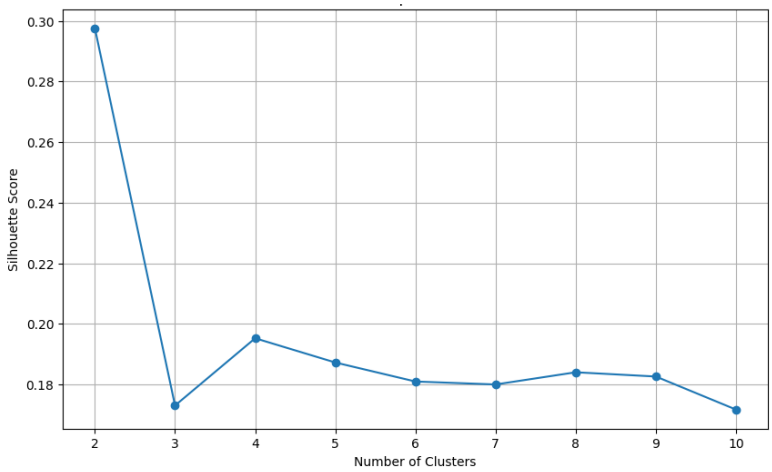

Figure 5.1 – Silhouette score for the grids using K-means.

To understand each cluster, it is important to observe their centroids (in the K-means algorithm, each centroid is defined as the average of all points belonging to that cluster). Figure 5.2 represents the centroids.

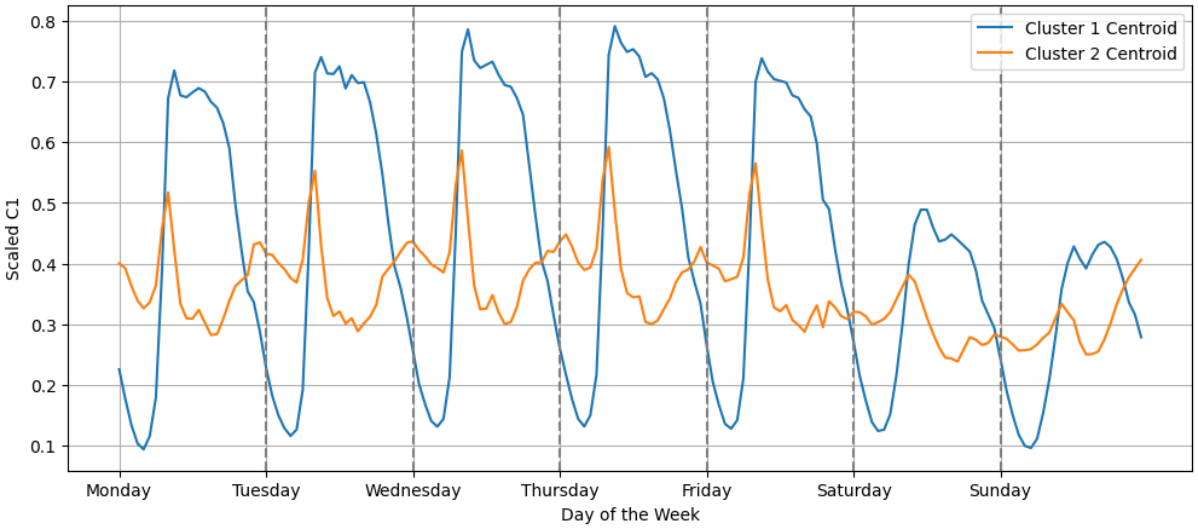

Figure 5.2 – Centroids of clusters by C1 time series using K-means.

The figure above represents two time series (one for each cluster), starting from Monday at 0:00 and finishing on Sunday at 23:00. For Cluster 1, the values are lower during nighttime, with a significant increase in the morning, remaining higher during the day and starting to decrease as the night progresses. This behavior is typical of "commercial" zones, where users commute for work. Activity is higher on weekdays compared to weekends and remains mostly constant during work hours.

Cluster 2 exhibits the opposite behavior. Values tend to be higher early in the morning and decrease to a lower value throughout the day, with another increase at night. The values also show smaller variance compared to Cluster 1. This behavior can represent either a "residential" zone or a "transition" zone, such as areas near highways, bridges, and airports.

The map of Lisbon representing the clusters can be seen in Figure 5.3. Based on the cluster centroids, Cluster 1 is labeled as "commercial" and Cluster 2 as "residential". As depicted, most of the city is clustered as commercial, especially the city center. This is explained by the concentration of jobs around the city center, including historical and tourist spots. Residential areas are typically located on the outskirts of the city, extending into other cities within the metropolitan area of Lisbon. This pattern aligns with commuting behaviors, where housing prices tend to be lower farther from the city center.

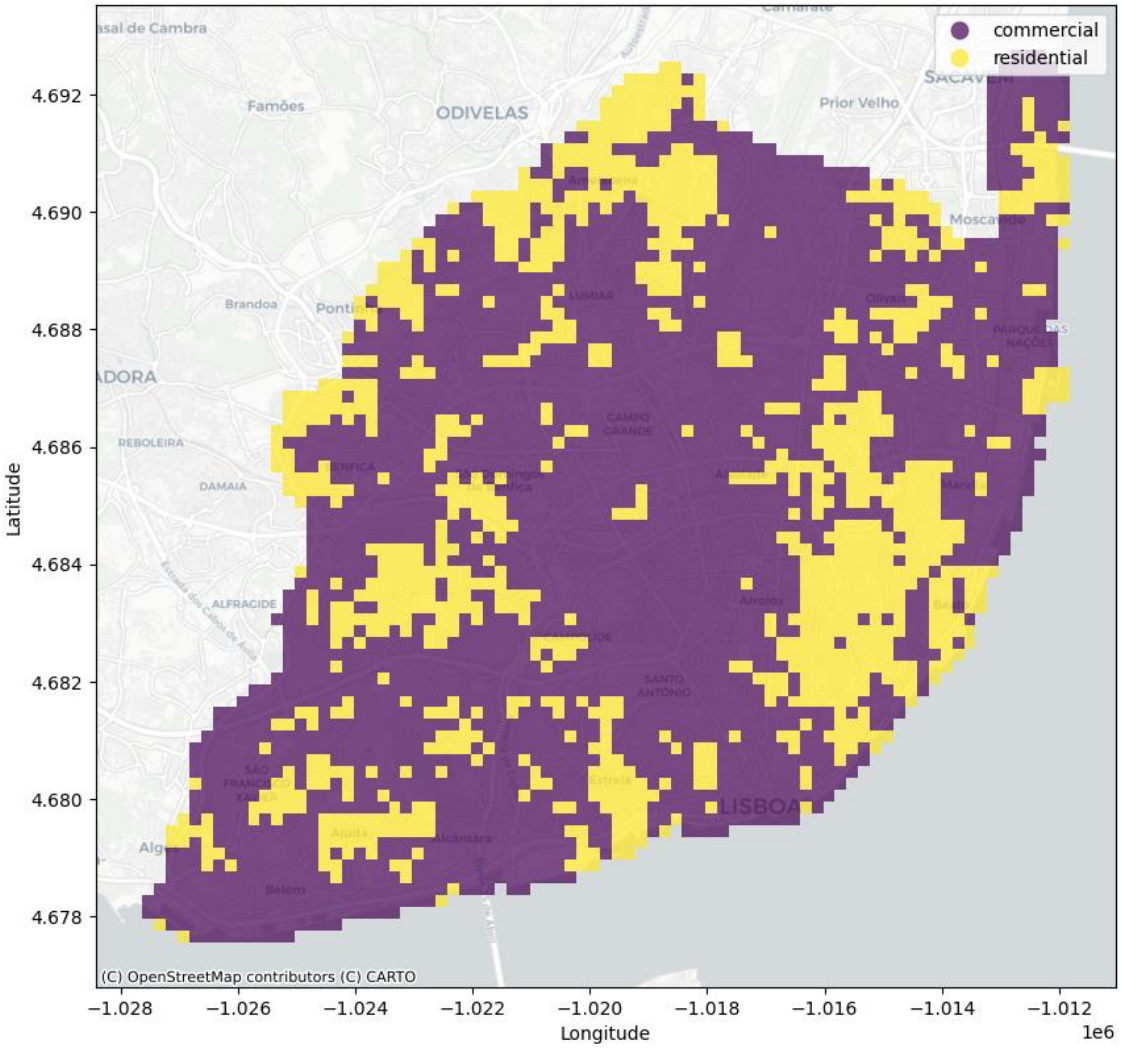

Figure 5.3 – Lisbon map clustered by C1 time series using K-means.

5.1.2. Hierarchical

Similar to the previous analysis, the study decided to cluster the grids using the time series values, now considering the hierarchical method. This method groups grids one by one, based on the distance between the points (the time series representation), as explained in Topic 4.4.1 and Figure 4.2.

The grids are grouped based on the distance between their time series, up until there is only one group. This process of joining the grids one by one can be illustrated in a dendrogram, as shown in Figure 5.4. To define the number of clusters in this method, a distance threshold is chosen to cross the dendrogram and split the clusters. Typically, the threshold is set at the largest gap that joins two different groups. From the dendrogram, the threshold was chosen at 50, since joining these two groups represented the largest distance, and also to keep the same number of clusters as in the previous method.

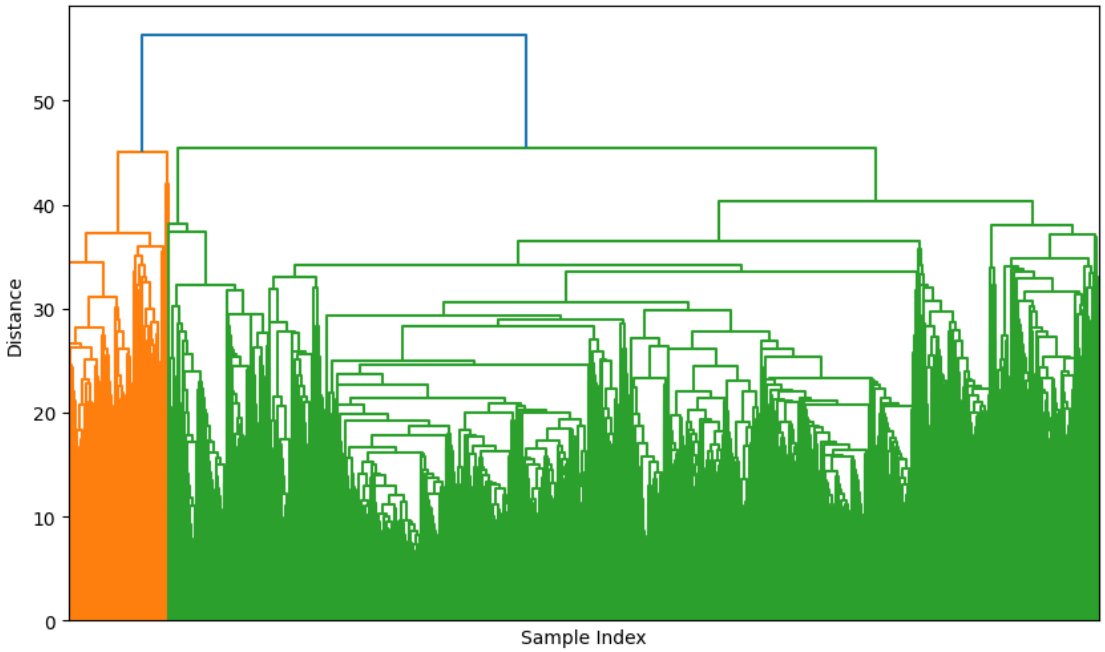

Figure 5.4 – Dendrogram of Hierarchical Clustering the grids time series.

From the clustering performed, the next step is to compare the cluster centroids from each method. This comparison is represented in Figure 5.5.

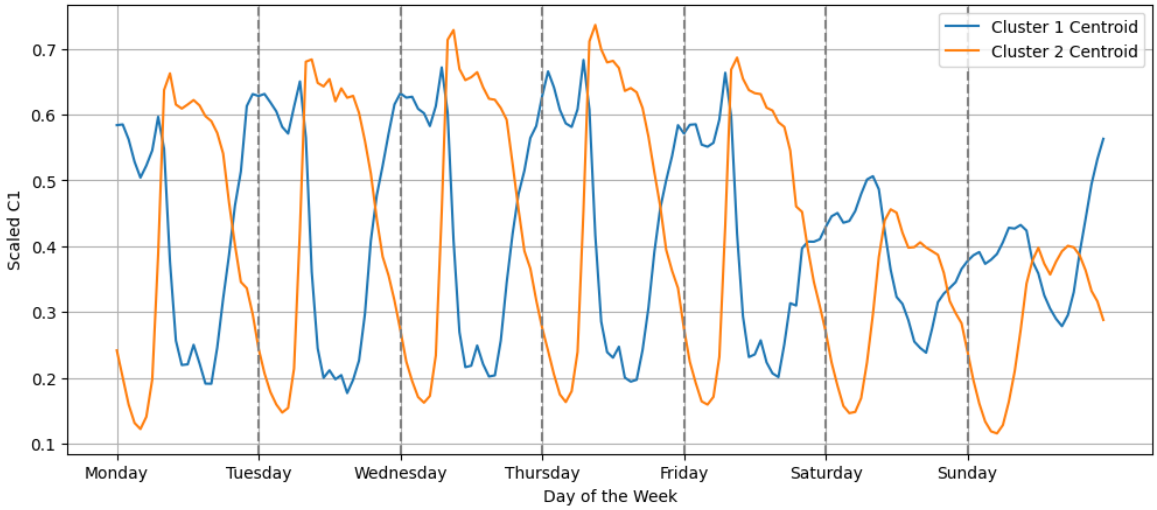

Figure 5.5 – Centroids of clusters by C1 time series using hierarchical clusters.

Like clustering by K-means, the hierarchical clustering also identifies Cluster 2 with a “commercial” behavior and Cluster 1 with a “residential” behavior. In this method, the residential cluster exhibits higher variance in its centroid throughout the day. Based on the map in Figure 5.6, this can be explained by the fact that there are fewer points in this cluster, leading to higher consistency across the grids, as they are closer to the scaled values from zero to one.

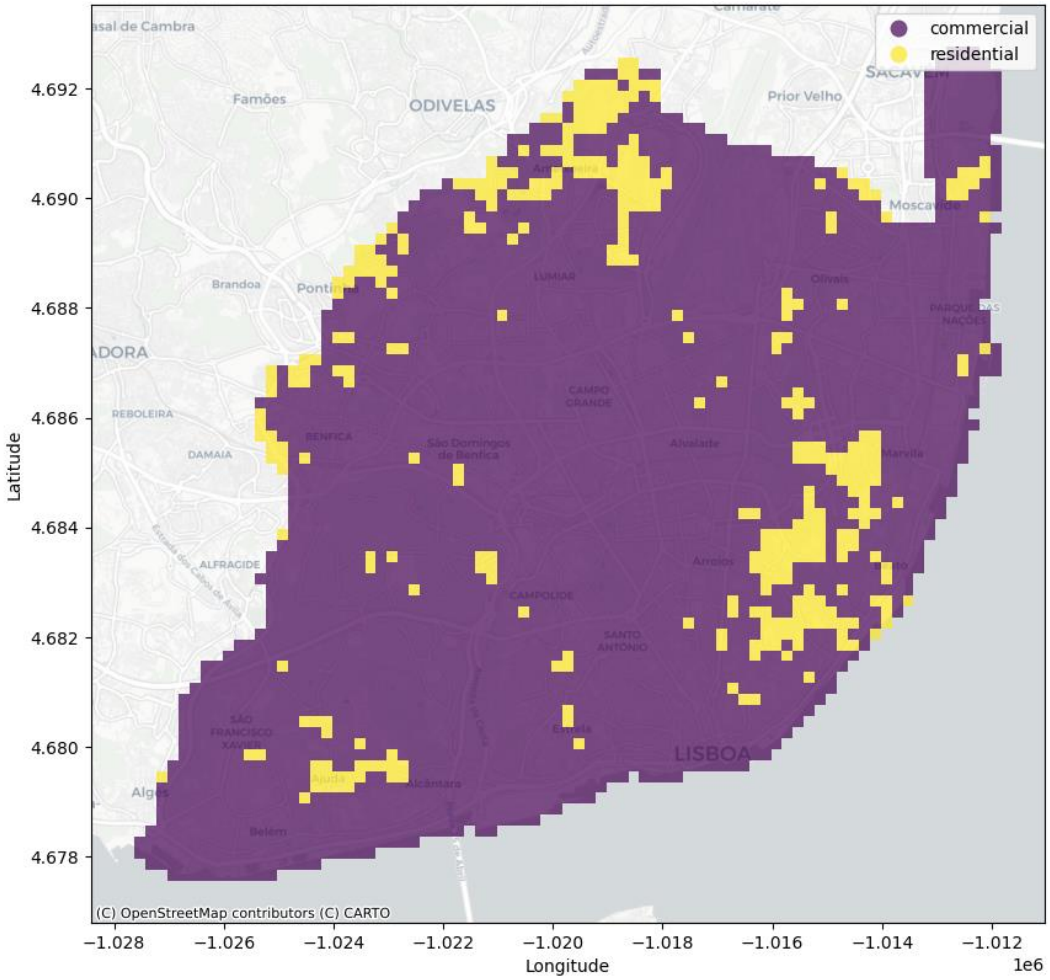

Figure 5.6 – Lisbon map clustered by C1 time series using hierarchical clustering.

When comparing the silhouette score from the K-means method to the Hierarchical method, the latter performed slightly better, with a score of 0.36041 against the score of 0.29754 from K-means. However, when comparing the map, most of Lisbon is classified as "commercial." Even though jobs grow around the city center and residences tend to expand into neighboring cities, it is unexpected that most of the city is classified as commercial.

One possible explanation for this classification is that when people reach home, they often switch from mobile data to wireless connections. This change could create a gap in tracking. The increase in the number of people working from home, especially in the period in analysis between 2021 and 2022, could also contribute to this behavior. Despite these factors, this classification provided the best results and will be discussed further in section 6.

5.2. CLUSTERING BY FEATURES

Exploring the features dataset allows for greater flexibility in analyzing the different characteristics of the data. However, this approach has the disadvantage of losing the temporal aspect. Therefore, a study must be conducted to better understand the available features and how they align with the main goals of the project.

As discussed in topic 4.3.1, the features selected for analysis were C1, C2, C11, E8, and E9. Figure 5.7 shows a heatmap illustrating the average values of these features across the week.

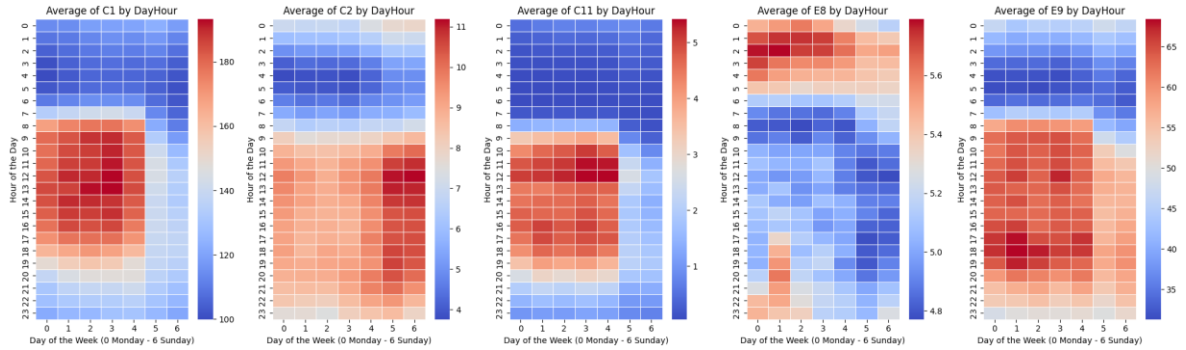

Figure 5.7 – Heatmap of features across the week.

As expected, C1, C11, and E9 are highly concentrated on weekdays during working hours, as this is when most people are at work and remain active on their devices. C2, which represents devices on roaming, also shows an expected pattern with higher values on weekends and late at night, as tourists tend to travel on weekends and stay out late. One unexpected result was E8, which shows higher concentrations during sleep times on weekdays.

This data analysis is important for defining which features would be included in the final features dataset. The creation of this dataset was explained in topic 4.3.4.

5.2.1. K-means

Once the features dataset was generated, the clustering process could begin. The first step, as before, is to define the number of clusters. For that, the silhouette score was used. As seen in Figure 5.8, the best number of clusters is 2.

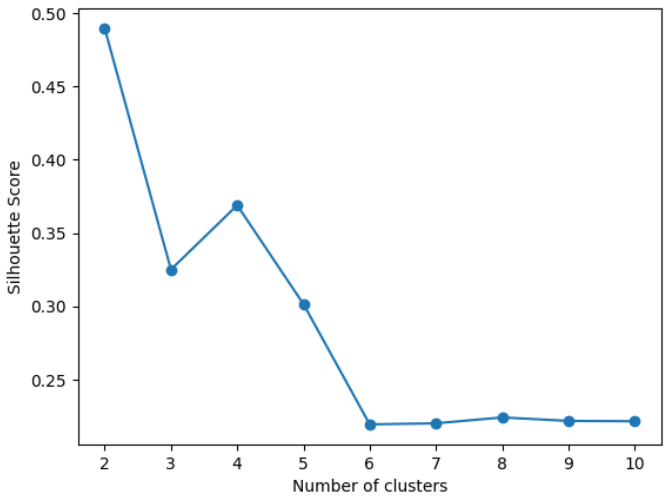

Figure 5.8 – Silhouette Score to define number of clusters K-means by features.

Table 5.1 – Centroids of clustering by features using K-means.

Feature	Cluster 1	Cluster 2
Avg_C1	0.060141	0.278318
Avg_C2	0.010609	0.119769
Avg_C11	0.034333	0.215647
Avg_E8	0.233281	0.294853
Avg_E9	0.179394	0.522304
Avg_day_peak	0.059394	0.258369
Max_diff_hour	0.477399	0.428792
Min_diff_hour	0.535004	0.568983
Avg_C2_C1_ratio	0.107004	0.257299
Weekday_C1	0.053973	0.252524
Weekday_C2	0.010535	0.119874
Weekend_C1	0.066072	0.295816
Weekend_C2	0.010779	0.119544
Working_C1_mean	0.043797	0.219275
Working_C2_mean	0.011028	0.106938

As observed from the table, cluster 2 corresponds to the most active zones in Lisbon. Every feature chosen for the study has a higher average value for cluster 2 compared to cluster 1. Because of this, cluster 1 will be defined as the "low activity" zone and cluster 2 as the "high activity" zone for this part of the study. The map of these clusters can be seen in Figure 5.9.

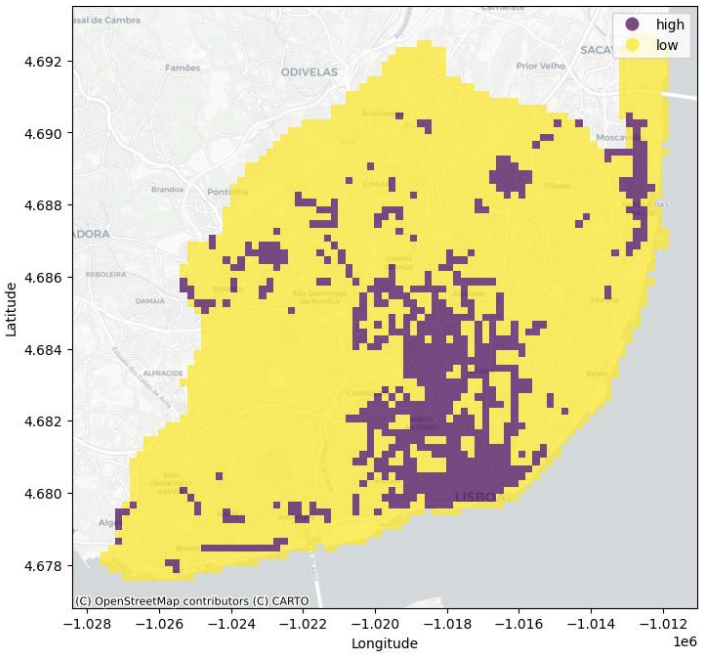

Figure 5.9 – Lisbon map clustered by features using K-means.

5.2.2. Hierarchical

For Hierarchical Clustering, the first step is to calculate the dendrogram. From Figure 5.10, it was decided to cluster into 2 groups as well, dividing at a distance of 15.

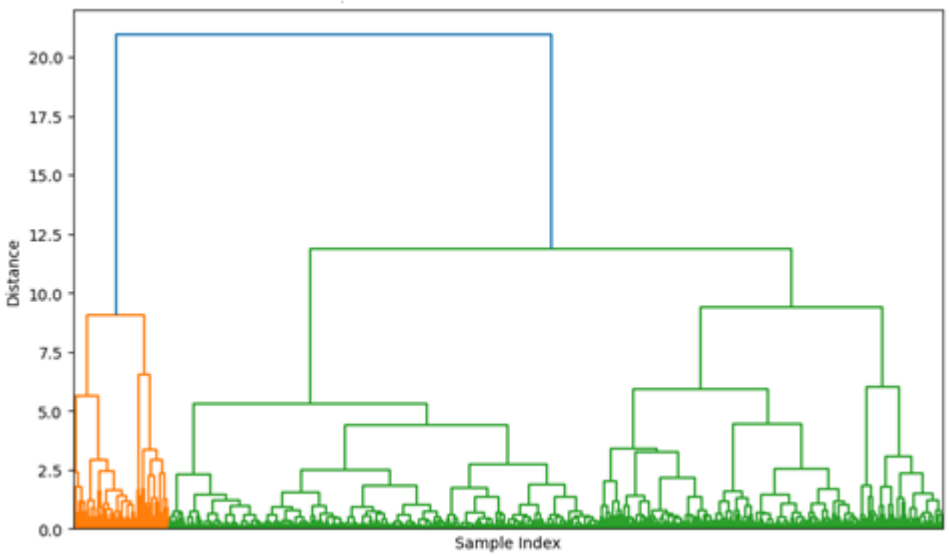

Figure 5.10 - Dendrogram of Hierarchical Clustering the grids features.

Table 5.2 – Centroids of clustering by features using Hierarchical Clustering.

Feature	Cluster 1	Cluster 2
Avg_C1	0.326265	0.071371
Avg_C2	0.161766	0.014001
Avg_C11	0.249657	0.044388
Avg_E8	0.298517	0.237671
Avg_E9	0.575917	0.199734
Avg_day_peak	0.300395	0.069846
Max_diff_hour	0.448302	0.471158
Min_diff_hour	0.523075	0.543353
Avg_C2_C1_ratio	0.312557	0.111991
Weekday_C1	0.294445	0.064405
Weekday_C2	0.161582	0.013977
Weekend_C1	0.352475	0.077135
Weekend_C2	0.162179	0.014061
Working_C1_mean	0.253820	0.053326
Working_C2_mean	0.141269	0.014325

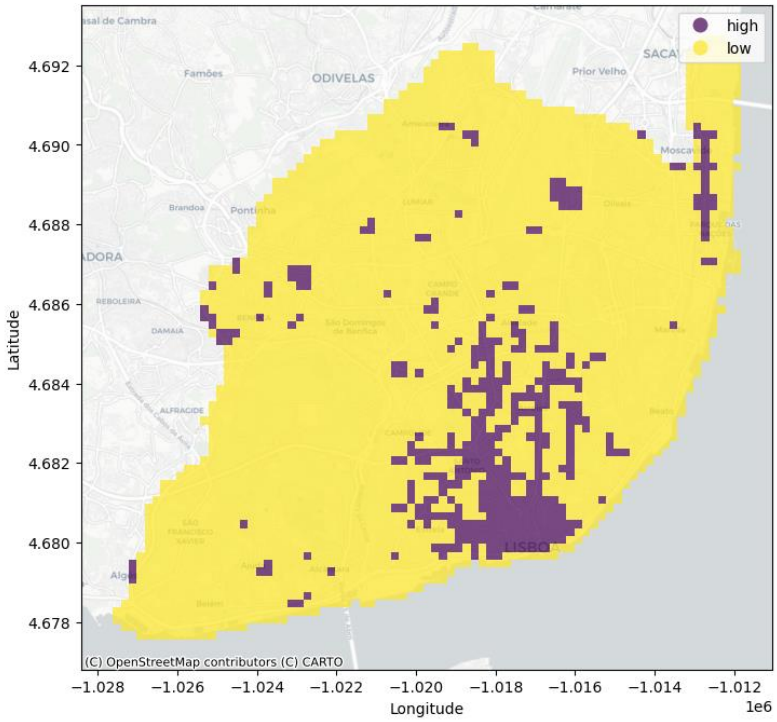

Figure 5.11 – Lisbon map clustered by features using Hierarchical Clustering.

The results obtained are similar to the previous method. Cluster 1 has higher average values than Cluster 2 (as seen in Table 5.2), categorizing Cluster 1 as "high" and Cluster 2 as "low". From Figure 5.11, the map of Lisbon shows a pattern like that of the K-means method. When comparing the silhouette scores of both methods, once again hierarchical clustering achieved a higher score of 0.70064 compared to 0.55767 from K-Means.

First thing to notice is that the features method has a significantly higher silhouette score than the time series method. One possible explanation is that the features method operates with lower dimensionality compared to the time series method, which encompasses 168 dimensions (7 days a week times 24 hours a day). Although 15 features were used, there is potential to further optimize by selecting fewer specific features that perform better.

However, the features method primarily distinguished grids with higher activity from those with lower activity, while the time series method provides a more detailed representation of grid performance over a week. Therefore, each clustering technique has its own advantages and disadvantages, and the choice should align with the study's objectives.

Lastly, the results obtained are not contradictory but complementary. While the time series method identified clusters as "commercial" and "residential" areas, the features method categorized clusters as "high" and "low" activity. By combining these results, it is possible to identify four groups: commercial areas with high and low activity, and residential areas with high and low activity.

6. DISCUSSION

This study aimed to cluster the grids in the city of Lisbon to assist in urban planning efforts aimed at improving urban flow across the city. Two distinct approaches were employed: one focused on analyzing time series data over the number of devices in each grid across the week, while the other defined features for each grid to perform the clustering techniques.

One issue identified when comparing both approaches was the disparity in silhouette scores. Although the features-based method achieved notably higher scores, it struggled to reveal patterns across the week. Therefore, time series clustering can serve as a complementary method in this study, offering insights into different aspects of the data.

The study also employed two different clustering techniques: K-Means and Hierarchical. In both methods (by features and by time series), the hierarchical method outperformed K-Means. Two possible explanations for this are that K-Means is more sensitive to outliers and performs better on spherical clusters. Given the potential noise in readings from the mobile operator, which could introduce outliers, and the fact that the features may not exhibit a spherical shape, it was expected that hierarchical clustering would perform better.

When performing time series clustering, the study identified two major groups: “commercial” and “residential”. The first one had higher values across the day, during working hours, with a high increase in the early morning and a high decrease in the evening to late in the night. The second group had the opposite behavior, with higher values at night, a high decrease in the morning and increasing again in the evening.

The commercial group is the main area where people work. This area has an expected behavior, with a high increase in the morning and high decrease on the evening and lower activity on weekends. To avoid congestions, these are possible areas to invest on urban housing. Although it is the most expensive region in Lisbon, an investment on newer houses, increasing the number of residential buildings and number of apartments could impact with a lower price on rent and the population would be closer to their jobs, wasting less time on transportation. If unfeasible, other strategies would be to incentivize commercial zones farther from this region, so the jobs would be more spread across Lisbon and neighboring cities, improving urban flow.

On the other hand, the residential zone is usually on the borders of the city and public transportation stations, where there is huge flow early in the morning and later in the evening, with fewer activity across the day and afternoon. Those areas have a growth in potential, as currently they are mostly residential and there is little activity over them. An investment for companies to move to those regions could increase the impact on those, increasing the activity on them across the day and bringing the companies closer to residential areas, reducing urban traffic.

When applying this clustering technique, the "commercial" cluster covered a larger area compared to the "residential" one, especially evident in the hierarchical clustering. Despite the expectation for residential zones to expand in neighboring cities over Lisbon, this outcome was unexpected. Several hypotheses were considered to explain this. During the dataset's timeframe, from September 2021 to August 2022, COVID-19 social distancing measures may have influenced the results. The "commercial" behavior observed (increasing during the day and decreasing at night) could simply reflect individuals working from home. Another possibility is that many people switch to using Wi-Fi networks at home, which are not tracked by mobile operators for this study.

The other strategy applied on the study (clustering by features) divided the city into 2 clusters. Based on the results obtained, the clusters were divided between “high” and “low” activity zones. This method, however, have fewer information about the population distribution across the day, only the averages defined on the beginning of the study. Because of that, this strategy can be improved by selecting better features.

Since this division provide less information over the time, the only general information obtained is that one cluster have higher activity over the other. Because of that, possible strategies for urban planning are to invest on the areas of low activity. With an increase in regions with fewer activity, the urban flow is better distributed across the city, reducing congestion, and developing the city proportionally.

About the differences between both strategies, each one has its own advantages. Clustering by time series was relatively easier, as it only had to define a single feature for the approach and compare between the grids. However, this strategy is more time sensitive, being more vulnerable to missing data and possible reading errors. Because of that, the preprocessing for this strategy need to be done more carefully. Besides that, a large number of timestamps can increase significantly the computational resources for this clustering approach.

Clustering by features is more flexible. Different features can be selected according to the study's objectives. Hence, it is crucial to have a clear discussion on the points to improve in the city to help decide the best features for clustering. The computational demands of this method are lower and can facilitate real-time monitoring of the city.

Overall, the power of data must walk together the city planning. Mobile operators can provide that information to measure on real time the concentration zones on the city. Deploying alerts on any anomalies can help identify traffic accidents quickly. Zones that are usually not on the scope of analysis can be seen as a potential growth and increase the city revenue. The possibilities are many and the city of Lisbon tends to grow when accurately measuring the data available.

7. CONCLUSIONS AND FUTURE WORKS

This study has discussed how urban planning can utilize mobile phone data as a useful resource to detect population patterns and help develop strategies to improve the city. A case study was applied to the city of Lisbon, retrieving data from a major mobile operator over a one-year period. The available data was analyzed to determine which features were most significant in identifying mobility patterns among the population.

The goal of the study was to divide the city into zones of similar behavior. Clustering techniques were applied to the dataset using two different approaches: clustering by time series and clustering by features. Both strategies provided different insights that could be used by urban planning to improve Lisbon.

One of the primary limitations of the study was related to the dataset itself, which consisted of over 60 GB of data, requiring significant computational resources for processing. As a result, a considerable portion of the dataset had to be reduced to facilitate the necessary operations, thereby limiting the achieved outcomes.

Regarding the data available, there were some issues with the quality of the data. There was a significant amount of missing data over the period of analysis. While small interruptions in the time series could easily be filled using certain techniques, large periods were more susceptible to errors, such as a large gap between February 8th and March 18th.

Another significant limitation was the identification of outliers across multiple grids exhibiting different behaviors. It was challenging to determine which data points constituted outliers. Ideally, time series analysis should have been utilized to decompose the grids into trend and seasonal components, identifying and mitigating residuals from the study.

During the development of the features dataset, challenges emerged in determining which features should be generated and selected for clustering techniques. With numerous features available and various operations possible across the time series data, defining key features proved to be complex.

Finally, the study limited itself to two simpler approaches when considering clustering, using k-means and hierarchical clustering. While other techniques, such as Self-Organizing Maps (SOMs), can provide deeper analysis of the data, time and computational resource constraints made the study remain on those two simpler approaches.

For future studies, developing regression models for each cluster to predict real-time mobile data values and detect anomalies could prove beneficial. This approach could enable quicker implementation of plans to enhance mobility, such as rerouting during traffic congestion or increasing public transportation services during events like concerts.

Additionally, text data, such as the most frequently used applications by users and the origin countries from roaming data, remains untapped. Analyzing user applications could foster partnerships with companies, improving targeted commercial advertisements and generating revenue for the city. Understanding the origin countries from roaming data could inform strategies for immigration services and tourism departments, aiming to attract tourists and manage immigration flows effectively.

In conclusion, the use of data is an important tool for the city planning and it can be used on many different strategies. It is an open field of possibilities, and the city of Lisbon tends to improve if the data is correctly analyzed and developed accurate plans from the results.

BIBLIOGRAPHICAL REFERENCES

Aghabozorgi, S., Seyed Shirkhorshidi, A., & Ying Wah, T. (2015). Time-series clustering – A decade review. Information Systems, 53, 16–38. https://doi.org/10.1016/j.is.2015.04.007

Aguiléra, A., Wenglenski, S., & Proulhac, L. (2009). Employment suburbanisation, reverse commuting and travel behaviour by residents of the central city in the Paris metropolitan area. Transportation Research Part A: Policy and Practice, 43(7), 685–691. https://doi.org/10.1016/j.tra.2009.06.004

Al Bassam, N., Hussain, S. A., Al Qaraghuli, A., Khan, J., Sumesh, E. P., & Lavanya, V. (2021). IoT based wearable device to monitor the signs of quarantined remote patients of COVID-19. Informatics in Medicine Unlocked, 24, 100588. https://doi.org/10.1016/j.imu.2021.100588

Alam, I., Farid, D. Md., & Rossetti, R. J. F. (2019). The Prediction of Traffic Flow with Regression Analysis. Em A. Abraham, P. Dutta, J. K. Mandal, A. Bhattacharya, & S. Dutta (Orgs.), Emerging Technologies in Data Mining and Information Security (p. 661–671). Springer. https://doi.org/10.1007/978-981-13-1498-8_58

Anagnostopoulos, T. (2021). A Predictive Vehicle Ride Sharing Recommendation System for Smart Cities Commuting. Smart Cities, 4(1), Artigo 1. https://doi.org/10.3390/smartcities4010010

Anas, A., Arnott, R., & Small, K. A. (1998). Urban Spatial Structure. Journal of Economic Literature, 36(3), 1426–1464.

Becker, R. A., Caceres, R., Hanson, K., Loh, J. M., Urbanek, S., Varshavsky, A., & Volinsky, C. (2011). A Tale of One City: Using Cellular Network Data for Urban Planning. IEEE Pervasive Computing, 10(4), 18–26. IEEE Pervasive Computing. https://doi.org/10.1109/MPRV.2011.44

Chidean, M. I., Jiménez Gil, L. I., Carmona-Murillo, J., & Cortés-Polo, D. (2023). Information theory based clustering of cellular network usage data for the identification of representative urban areas. Digital Communications and Networks. https://doi.org/10.1016/j.dcan.2023.07.002

Demissie, M. G., de Almeida Correia, G. H., & Bento, C. (2013). Intelligent road traffic status detection system through cellular networks handover information: An exploratory study. Transportation Research Part C: Emerging Technologies, 32, 76–88. https://doi.org/10.1016/j.trc.2013.03.010

fmarthoz. (2021, julho 26). Feature selection for K-means. Analytics Vidhya. https://medium.com/analytics-vidhya/k-means-algorithm-in-4-parts-4-4-42bc6c781e46

Frias-Martinez, V., Soguero, C., & Frias-Martinez, E. (2012). Estimation of urban commuting patterns using cellphone network data. Proceedings of the ACM SIGKDD International Workshop on Urban Computing, 9–16. https://doi.org/10.1145/2346496.2346499

Ghahramani, M., Zhou, M., & Wang, G. (2020). Urban sensing based on mobile phone data: Approaches, applications, and challenges. IEEE/CAA Journal of Automatica Sinica, 7(3), 627–637. IEEE/CAA Journal of Automatica Sinica. https://doi.org/10.1109/JAS.2020.1003120

Heka.ai. (2022, junho 9). Time series clustering. Medium. https://heka-ai.medium.com/time-series-clustering-b84bcaaa63ac

Herrera, J. C., Work, D. B., Herring, R., Ban, X. (Jeff), Jacobson, Q., & Bayen, A. M. (2010). Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment. Transportation Research Part C: Emerging Technologies, 18(4), 568–583. https://doi.org/10.1016/j.trc.2009.10.006

Jahani, E., Sundsøy, P., Bjelland, J., Bengtsson, L., Pentland, A. 'Sandy', & de Montjoye, Y.-A. (2017). Improving official statistics in emerging markets using machine learning and mobile phone data. EPJ Data Science, 6(1), Artigo 1. https://doi.org/10.1140/epjds/s13688-017-0099-3

Jiang, S., Ferreira, J., & González, M. C. (2012). Clustering daily patterns of human activities in the city. Data Mining and Knowledge Discovery, 25(3), 478–510. https://doi.org/10.1007/s10618-012-0264-z

Karikoski, J., & Soikkeli, T. (2013). Contextual usage patterns in smartphone communication services. Personal and Ubiquitous Computing, 17(3), 491–502. https://doi.org/10.1007/s00779-011-0503-0

Kung, K. S., Greco, K., Sobolevsky, S., & Ratti, C. (2014). Exploring Universal Patterns in Human Home-Work Commuting from Mobile Phone Data. PLOS ONE, 9(6), e96180. https://doi.org/10.1371/journal.pone.0096180

Kyte, M., Khan, A., & Kagolanu, K. (1993). Using Machine Vision (Video Imaging) Technology To Collect Transportation Data. Transportation Research Record, 1412, 23–32.

Lawgic. (2022, agosto 7). Internet Protocol Detail Records IPDR reveal a lot of secrets about WhatsApp, Telegram and other IP based Voice / Video Calling Services. Lawgic. https://lawgic.info/how-ipdr-internet-protocol-detail-records-can-reveal-a-lot-of-secrets-about-whatsapp-voice-calling/

Ma, X., Liu, C., Wen, H., Wang, Y., & Wu, Y.-J. (2017). Understanding commuting patterns using transit smart card data. Journal of Transport Geography, 58, 135–145. https://doi.org/10.1016/j.jtrangeo.2016.12.001

Mahmassani, H. S., Joseph, T., & Jou, R.-C. (1993). Survey approach for study of urban commuter choice dynamics. Transportation Research Record, 1412, 80–89.

Martin, D., Gale, C., Cockings, S., & Harfoot, A. (2018). Origin-destination geodemographics for analysis of travel to work flows. Computers, Environment and Urban Systems, 67, 68–79. https://doi.org/10.1016/j.compenvurbsys.2017.09.002

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M. J., & Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. IEEE Transactions on Knowledge and Data Engineering, 33(8), 3048–3061. IEEE Transactions on Knowledge and Data Engineering. https://doi.org/10.1109/TKDE.2019.2962680

Mokhtari, A., Ghorbani, N., & Bahrak, B. (2022). Aggregated Traffic Anomaly Detection Using Time Series Forecasting on Call Detail Records. Security and Communication Networks, 2022, e1182315. https://doi.org/10.1155/2022/1182315

Neuhaus, F. (2010). UrbanDiary—A Tracking Project: Capturing the beat and rhythm of the city: Using GPS devices to visualise individual and collective routines within Central London. 1(2).

Phithakkitnukoon, S., Sukhvibul, T., Demissie, M., Smoreda, Z., Natwichai, J., & Bento, C. (2017). Inferring social influence in transport mode choice using mobile phone data. EPJ Data Science, 6(1), Artigo 1. https://doi.org/10.1140/epjds/s13688-017-0108-6

Puri, C., Kooijman, G., Vanrumste, B., & Luca, S. (2022). Forecasting Time Series in Healthcare With Gaussian Processes and Dynamic Time Warping Based Subset Selection. IEEE journal of biomedical and health informatics, PP. https://doi.org/10.1109/jbhi.2022.3214343

Qiao, Y., Cheng, Y., Yang, J., Liu, J., & Kato, N. (2017). A Mobility Analytical Framework for Big Mobile Data in Densely Populated Area. IEEE Transactions on Vehicular Technology, 66(2), 1443–1455. IEEE Transactions on Vehicular Technology. https://doi.org/10.1109/TVT.2016.2553182

Sagl, G., Delmelle, E., & Delmelle, E. (2014). Mapping collective human activity in an urban environment based on mobile phone data. Cartography and Geographic Information Science, 41(3), 272–285. https://doi.org/10.1080/15230406.2014.888958

Steenbruggen, J., Tranos, E., & Nijkamp, P. (2015). Data from mobile phone operators: A tool for smarter cities? Telecommunications Policy, 39(3), 335–346. https://doi.org/10.1016/j.telpol.2014.04.001

Thuillier, E., Moalic, L., Lamrous, S., & Caminada, A. (2018). Clustering Weekly Patterns of Human Mobility Through Mobile Phone Data. IEEE Transactions on Mobile Computing, 17(4), 817–830. IEEE Transactions on Mobile Computing. https://doi.org/10.1109/TMC.2017.2742953

Xu, F., Lin, Y., Huang, J., Wu, D., Shi, H., Song, J., & Li, Y. (2016). Big Data Driven Mobile Traffic Understanding and Forecasting: A Time Series Approach. IEEE Transactions on Services Computing, 9(5), 796–805. IEEE Transactions on Services Computing. https://doi.org/10.1109/TSC.2016.2599878

Yang, T. ([s.d.]). Understanding commuting patterns and changes: Counterfactual analysis in a planning support framework—Tianren Yang, 2020. Recuperado 20 de novembro de 2023, de https://journals.sagepub.com/doi/full/10.1177/2399808320924433

Yu, Q., Li, W., Yang, D., & Zhang, H. (2020). Mobile Phone Data in Urban Commuting: A Network Community Detection-Based Framework to Unveil the Spatial Structure of Commuting Demand. Journal of Advanced Transportation, 2020, e8835981. https://doi.org/10.1155/2020/8835981

Zhang, P., Zhou, J., & Zhang, T. (2017). Quantifying and visualizing jobs-housing balance with big data: A case study of Shanghai. Cities, 66, 10–22. https://doi.org/10.1016/j.cities.2017.03.004

Zhang, Y. (2014). User mobility from the view of cellular data networks. IEEE INFOCOM 2014 - IEEE Conference on Computer Communications, 1348–1356. https://doi.org/10.1109/INFOCOM.2014.6848068

Zhao, P., Lu, B., & de Roo, G. (2011). The impact of urban growth on commuting patterns in a restructuring city: Evidence from Beijing. Papers in Regional Science, 90(4), 735–754. https://doi.org/10.1111/j.1435-5957.2010.00343.x

Zhou, C., Frankowski, D., Ludford, P., Shekhar, S., & Terveen, L. (2007). Discovering personally meaningful places: An interactive clustering approach. ACM Transactions on Information Systems, 25(3), 12-es. https://doi.org/10.1145/1247715.1247718

Zhu, J., & Fan, Y. (2018). Commute happiness in Xi'an, China: Effects of commute mode, duration, and frequency. Travel Behaviour and Society, 11, 43–51. https://doi.org/10.1016/j.tbs.2018.01.001

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa