Analyzing urban mobility based on smartphone data: the Lisbon case study

Daniel Leal[1], Vitória Albuquerque[2[0000-0001-9684-968X]], Miguel Sales Dias[1[0000-0003-3292-4454]] , João Carlos Ferreira[1,3[0000-0002-6662-0806]]

[1] Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, 1649-026 Lisboa, Portugal
[2] NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal
[3] INOV INESC Inovação—Instituto de Novas Tecnologias, 1000-029 Lisboa, Portugal
drllo@iscte-iul.pt

Abstract. Our paper addresses the mobility patterns in Lisbon in the vicinity of historical and transportation points of interest, with a case study conducted in the parish of Santa Maria Maior, a vibrant touristic neighborhood. We propose a data science-based approach to analyze such patterns. Our dataset includes five months of georeferenced mobile phone data, collected during late 2021 and early 2022, provided by the municipality of Lisbon. We performed a systematic literature review, using the PRISMA methodology and adopted the CRISP-DM methodology, to perform data curation, statistical and clustering analysis, and visualization, following the recommendations of the literature. For clustering we used the DBSCAN algorithm. We found eight clusters in Santa Maria Maior, with outstanding clusters along 28-E tram and Lisbon Cruise Terminal, where mobility is high, particularly for non-roaming travelers. This paper contributes to the digital transformation of Lisbon into a smart city, by improving improved understanding of urban mobility patterns.

Keywords: smartphone data, urban mobility, visualisation, point of interest, DBSCAN, PRISMA, CRISP-DM.

1 Introduction

1.1 Motivation and topic relevance

The analysis of available Internet of Things (IoT) data in urban settings by relevant stakeholders, shows that city decision-makers can alleviate urbanization's pressures by providing a new experience for citizens, making their day-to-day life more comfortable and secure. In smart cities, IoT refers to the use of smart computing and networking technology and linked devices for real-time data collection. Rising urbanization, increased demand for efficient infrastructure in metropolitan areas, as well as for energy-efficient resources, traffic management, waste management, public safety, and security, which in turn, are development factors for the total market. Connected inter-

net technologies and devices can be used to alleviate problems, improve the quality of residents' life, and minimize resource consumption in smart cities.

In urban settings, it has become increasingly crucial to determine the location of mobile phone users in the Global System for Mobile (GSM) networks. The location of a mobile phone can be determined using the network architecture of the telecom service provider. It is possible to collect raw radio data of a handset using the subscriber identity module (SIM) in GSM and Universal Mobile Telecommunications System (UMTS) devices. The precision of any localization system is critical to the success of the technology in the long run, and it is determined by the density of cellular base stations, with urban areas obtaining the best potential accuracy due to the increased number of cell towers, as well as the use of the most up-to-date timing methods and technologies. Numerous factors can affect the accuracy of location data, including its source, which may include Global Position System (GPS) signals, Wi-Fi, or cell tower triangulation.

Rush hours and traffic jams have become part of our daily routines over the years, as well as the research drive to reduce this phenomenon. As result, it is becoming increasingly vital to revolutionize traffic management in urban areas using data and a variety of computing methods to help cities to understand what is happening and provide new mobility strategies.

The availability of Vodafone Portugal [1] mobile phone data provided by Câmara Municipal de Lisboa (CML), opened an opportunity and interest to study this data in the scope of urban mobility in the city of Lisbon, especially to understand how, when and where people travel in the city. Considering this data, the aim of our research is to understand mobility patterns in Lisbon, by performing analysis and visualization of mobility phenomena during a given time period.

The results of this study will provide knowledge to the policy and decision makers at CML, enabling better mobility patterns understanding, as well as the implementation of sustainable urban mobility and tourism strategies for the city.

1.2 Research question and objective

This research theme was proposed by Iscte in partnership with CML's Center for Management and Urban Intelligence by the LxDataLab [2] and also in partnership with Vodafone Portugal.

Our research question can be stated in the following way: "what are the mobility patterns of smartphone users in the city of Lisbon related to points of interest, namely, historical places and public transportation?"

This research question led us to our research objective that, in short, aims to understand the mobility patterns in Lisbon by analysing mobile phone data, in the vicinity of the mentioned points of interest. We propose to perform analysis and visualization of mobile Vodafone data and open-source mapping data to identify mobility patterns in Lisbon, using data mining and visualization. Our data mining approach, adopts the CRISP-DM methodology [3], [4], and for the modelling, we will use statistical analysis and cluster analysis, this last one with the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) method, following literature recommendations.

1.3 Structure

This paper is organized in four sections. In section 1, we introduce the theme of the paper, the topic context, research questions and goals, methodology, and structure. Section 2 introduces the results of our systematic literature review and bibliometric analysis using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [5] with findings on the latest state-of-the-art methodologies applied to urban mobility behavior patterns, based on the analysis of mobile phone data. In section 3, we use a data science approach to perform data mining, namely, adopting statistical analysis, DBSCAN clustering and visualization. In section 4, we discuss our results and research limitations, present our conclusions, and propose future lines of research work.

2 Literature review

2.1 Methodology

PRISMA [5] was applied with the purpose of identifying, evaluating, and critically appraising research, to provide an answer to a well-formulated query related to our research question. This methodology is a set of elements for systematic reviews and meta-analyses that is scientific proof.

2.2 Results

In this study we used the following keywords to query academic repositories (Scopus and Web of Science): “data mining” and “machine learning” and “smartphone” and “data”. It returned the most relevant papers, that were screened for eligibility and full text reading, resulting in a total of 12 papers that were included in our bibliometric analysis.

Based on the articles resulting from the PRISMA survey, we analyzed their methods and applications. We observed a trend in the application of DBSCAN method [6]–[9]. Other methods include visualization and analysis of mobility patterns with and without point of interest (POI) [10]–[15], k-means clustering algorithm [16], Dynamic Time Warping (DTW) [12], and analytic methods such as Point Density and Kernel Density Estimation [17].

We identified the use of the DBSCAN method in "Vehicular traffic flow intensity detection and prediction through mobile data usage" [6] where its application is made in an artificial neural network trained with the traffic levels of the network nodes in a time series to predict the traffic of the nodes. In paper "A cluster-Based Approach Using Smartphone Data for Bike-Sharing Docking Stations Identification: Lisbon Case Study" [16] for the identification of soft mobility hotspots at specific bike share docking stations using k-means clustering algorithm. Moreover, "Spatio-Temporal Mining To Identify Potential Traffic Congestion Based On Transportation Mode" [7] for the identification of potential traffic congestion using DBSCAN clustering algo-

rithm; “Understanding individual mobility pattern and portrait depiction based on mobile phone data” [8] studied the application for individual mobility pattern analysis and portrayed the depiction in various Chinese cities. Also, "Clustering Large-Scale Origin-Destination Pairs: A Case Study for Public Transit in Beijing" [9]. study applied as well, DBSCAN to determine the bus passengers in Beijing mobility patterns.

We also identified visualizations and statistical analysis methods in the articles "Applying Big Data Analytics to Monitor Tourist Flow for the Scenic Area Operation Management" [10] in which these methods are applied to the identification of tourist movement in Beijing. "Understanding Human Mobility Flows from Aggregated Mobile Phone Data" [11] used these methods to identify the population behavior in Milan. "Extracting Dynamic Urban Mobility Patterns Phone Data" [12] also used these methods to identify urban mobility patterns. The research "Ensemble-spotting: Ranking urban vibrancy via POI embedding with multi-view spatial graphs" [13] study identified mobility patterns with POIs to the discover the association between vibrant communities and geographical items. The research "Using bundling to visualize multivariate urban mobility structure patterns in the São Paulo Metropolitan Area"[14] identified spatial grouping and some visualization using the application of bundling approach to support multi-attribute trail datasets in the São Paulo metropolitan area.

The identification of urban mobility patterns in the city of Shanghai used an analytical approach with Point Density and Kernel Density Estimation in "Role of big data in development of smart city by studying the density of citizens in Shanghai" [17] also to be considered.

3 Data Mining

The data mining CRISP-DM methodology [3], [4] is implemented throughout this research.

3.1 Business Understanding

LxDataLab [2] is supported by CML and was established to respond to the need to build analytical solutions for the city of Lisbon, capable of enhancing urban planning, and improve resilience, security, mobility, operational, and emergency management in the city, using innovative data analysis and machine learning techniques.

LxDataLab launched yearly challenges to the academia and research communities to understand different city domains: environment, energy, citizen, economy, governance, mobility, and quality of life.

This study addresses challenge 70 theme on “Mobility in the city of Lisbon based on mobile phone data”. This challenge in the urban mobility domain, in collaboration with a mobile service operator (Vodafone), aims to understand how people handling a mobile phone move in the city, which is fully in line with our research objectives.

This paper tackles this challenge by analyzing the georeferenced data collected by Vodafone during a five-month period, from September 2021 to January 2022 and answering to our research question. In essence, the study aims to build an analytical research model centered on the CML smart city framework, looking at the mobility

patterns of smartphone users (nationals or roaming users), looking particularly at points of interest in the city, namely historic places and public transportation, helping decision makers of CML in the area of urban mobility.

3.2 Data Understanding

Five datasets were provided for each month, namely, September, October, November, December 2020, and January 2021. The data was compiled into 3,743 200-by-200 square meters (a grid of quadrants or quads).

According to Vodafone's metadata, there were no records reported with values less than 10 devices, and data was gathered every five minutes. Each monthly dataset provides the number of devices present in a certain quad every 5 minutes (along with a time marker), or more than 5 minutes for roaming and non-roaming, city enters and exits, terminal exits from the quad, top ten roaming nations and top ten applications, and downstream and upstream rates.

We have seventeen million records in the September dataset (17,233,318), thirty-two million records in the October dataset (32,627,308), twenty-one million records in the November dataset (21,619,292), thirty-three million records in the December dataset (33,121,657), and thirty-three million records in the January dataset (33,344,624). This resulted in a cumulative total of roughly 137 million records spread across five months.

An additional dataset with geoinformation, known as Vodafone grid, was provided by CML, and combined with the monthly datasets. This dataset complements the monthly datasets by containing information regarding the parish, street name, neighborhood or zone, position, and geometric information of the squares. It should be noted that two columns are shown for Lisbon parishes (freguesia and freguesias), which differ due to parish renaming and merging since November 8, 2012 [18] . As such we used the updated parishes information, set up after 2012.

3.3 Data Preparation

We included in our data type information, three ordinal qualitative variables - extract_year_2, extract_month_3, and extract_day_4 - and three continuous variables – Grid_ID, Datetime, and C3 or C.

Some of the datasets (September, November, December, and January) contain 44 nulls in column C3/C4, 30 nulls in column extract_year_2, and 43 nulls in column extract_day_4. We eliminated their entries due to the small quantity of nulls in the datasets.

After cleaning, we retained the following number of records: sixteen million records in the September dataset (16,166,066), thirty million records in the October dataset (30,604,296), twenty million records in the November dataset (20,142,789), thirteen million records in the December dataset (13,048,266), and thirty-one million records in the January dataset thirty-one million (31,277,197). This resulted in a total 111 million records to be used in this research, meaning that nearly 26 million records were deleted. The listing of the column "nome" values was visually analyzed to remove highways from the datasets, as these locations are prone to congestion, leading

us to misinterpretation of the data and misconception of our objective. Therefore, the following road routes and were removed from the "name" column: "A5", "Eixo Norte-Sul", "CRIL", "2ª Circular", and "A2", given that they correspond to is arriving, leaving or crossing the city.

A Python script was developed to group the data for a given month by parish and add up each parish's number of devices. The result of this analysis determined which Lisbon parish was going to be selected to be our case study. Additionally, we analyzed and visualized the number of stopped devices, the number of historical Points-of-Interest - POIs, the number of bus stop POIs, the number of metro station POIs, and the number of train station POIs, in the parishes of Lisbon, using choropleth maps. These POIs are related with tourism and sightseeing themes, arising from the combination of transportation and historical landmarks, that were chosen for this analysis to understand how people move in the city and how this behavior is related with the mentioned POIs, and to narrow our data modeling study to an outstanding parish.

POIs data was collected from the OSMnx library [19] via category-specific queries. We created two queries (due to museums' inclusion in the tourism category) for the historical POIs. For the train transportation POIs, a search for train stations was conducted, but the results also contained metro POIs, which were subsequently merged. Finally, for collecting bus stops and metro stations, a direct and simple search was sufficient.

In December (see Fig. 1), the non-roaming map shows that the interior and north parishes of Lisbon are more likely to have more devices. Avenidas Novas and Alvalade, for example, have more than 310 million devices, probably due to the traffic, workplaces, universities and cultural places location. On the other hand, the roaming map shows that the inner core of Lisbon, from Avenidas Novas to Santa Maria Maior, has more devices. Santa Maria Maior parish has more than 26 million devices, followed by Misericórdia, Santa António, Avenidas Novas, and Olivais, with more than 13 million. Estrela and Arroios have more than 7 million, and the remaining parishes have fewer than 7 million devices.

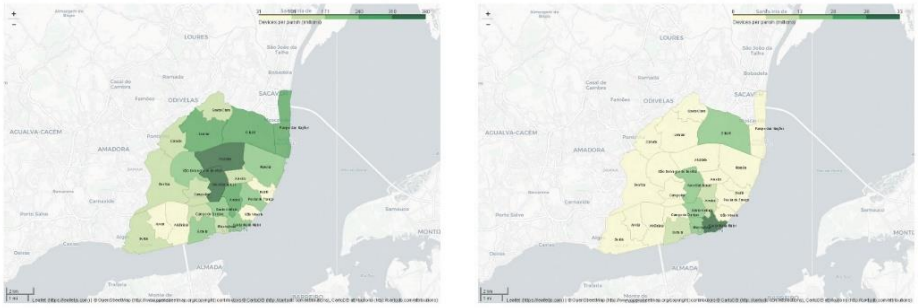

Fig. 1. Average number of devices (millions) per Lisbon parish in December: non-roaming (left) and roaming (right). Dark green areas have a higher number of smartphone devices.

Santa Maria Maior is the parish with the higher number of historical POIs, and we decided this criterion to analyze further this parish as a case study in our paper. In this

scenario, we recognized that Santa Maria Maior has not the higher number of observed devices but, still holds many data observations (between 31 million and 101 million).

3.4 Modeling

In this section, we present the data model results regarding the Santa Maria Maior parish, by analyzing mobile phone data and POIs data, showing insights on people's mobility patterns in this parish. We analyzed all the months of the datasets although, for the purpose of this paper, we are only presenting the month of December.

3.4.1 Statistical Analysis Model

For our statistical analysis, we started by analyzing POIs data, followed by mobile phone data and a combined analysis of both. Finally, we applied the DBSCAN clustering algorithm to the datasets, given that it is a technique adopted by the literature to similar problems, as shown in our literature survey.

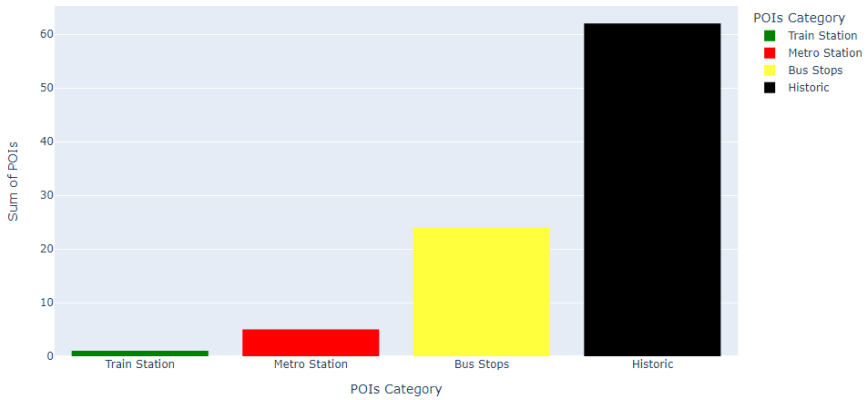

Fig. 2. POIs category histogram.

Fig. 2 shows a histogram depicting the total number of POIs, where colors of the public transportation-related histogram categories, such as metro station, railway station, and bus stop, were chosen based on the colors of their respective logos. In **Fig. 2**, we can observe that historical points of interest shows the highest number of POIs (62), followed by bus stops (24), metro stations (5) and train stations (1).

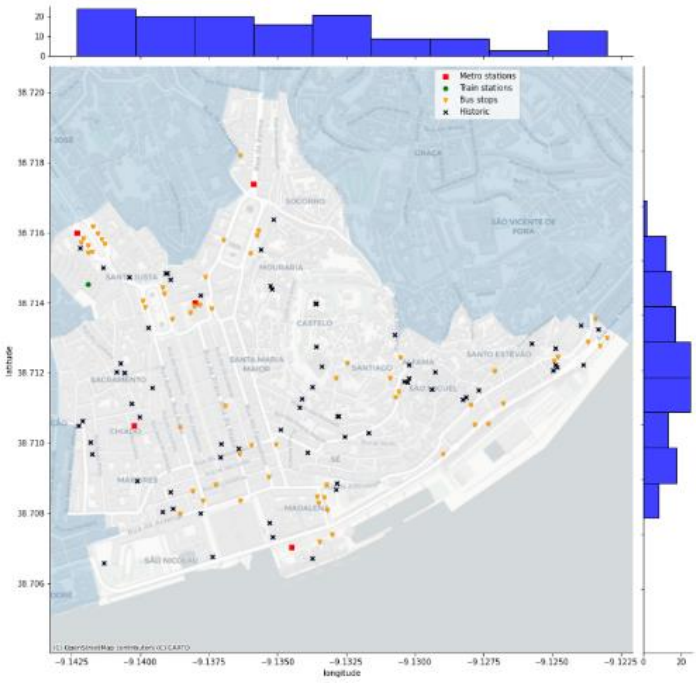

Fig. 3. POIs location distribution in Santa Maria Maior.

Fig. 3, depicts Santa Maria Maior parish POI categories locations (historical, bus stops, metro stations, and train stations), including two bar charts, of longitude (x-axis) and latitude (y-axis). Regarding the public transportation POIs, we can notice the multimodality nature of the system, particularly in Rossio, in the Santa Justa area, where there are 9 bus stops, 1 metro station, and 1 train station. This bridges mobility connections both within and outside this parish. The presence of bus stops and historic points of interest across the parish of Santa Maria Maior is noticable. We can observe the maximum concentration of POIs at the latitude between 38.710 and 38.714 with slightly over 40 points covered in total, with the distribution of POIs becoming increasingly smaller as one proceeds away from these locations. These correspond to stops and historical locations, and metro station serving the areas of Santo Estêvão, São Miguel, Alfama, Santiago, Sacramento, Chiado, and Castelo. The highest concentration in longitude is between -9.1425 and -9.1400, with about 20 points covered, and these POIs are largely bus stops and historical places, followed by metro stations and the train station, that spread in the areas of Santa Justa, Sacramento, Chiado, Mártires, and São Nicolau.

In our our data model and for each month, we associated the number of devices present in a given square polygon (of the 200mx200m grid of the data resolution), with the various POIs categories considered for this work, with the following technique: if a given POI point is inside a given square polygon, that POI will receive the number of devices collected for that polygon. Thus, we can conclude that *x* number of devices remained stationary for more than five minutes in a specific POI. The limita-

tion of our approach is that different POIs, regardless of category, have the same number of devices associated, if they belong to the same polygon.

Fig. 4. Historic POIs (red dots), included in smartphone data polygons in December 2021, in Santa Maria Maior: non-roaming (left) and roaming (right).

In **Fig. 4**, we can notice that, for both non-roaming and roaming data, it is highly concentrated in the Chiado area. We can observe also some orphan POIs, i.e., points that will not be considered since they were located beyond the data polygons associated with the parish in study.

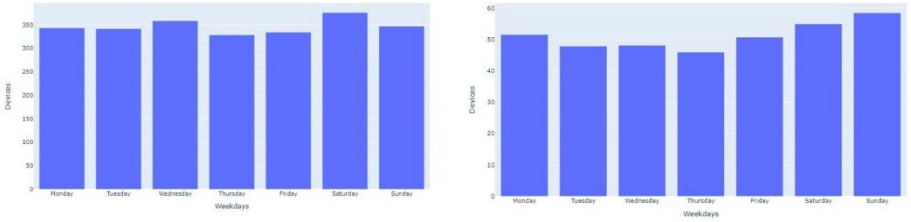

Fig. 5. Average thousand devices in December weekdays in Santa Maria Maior: non-roaming (left) and roaming (right).

Fig. 5, depicts two plot bars with average number of non-roaming and roaming devices, for the seven days of the week.

Observing the non-roaming plot bar, the weekly behavior on the number of devices tended to be very similar from Monday to Tuesday, with almost 350 thousand devices, rising on Wednesday with over 350 thousand devices, dropping to a maximum minimum of around 325 thousand devices on Thursday, and starting to rise until Saturday, when it reaches its highest value, around 375 thousand devices, and then dropping on Sunday. We observed that all weekdays have values close to 350 thousand devices, except for Wednesday, which has slightly more than 350 thousand devices.

The graph for the roaming showed that the weekly behavior tended to decrease from Monday with just over 50 thousand devices, followed by Tuesday and Wednesday with under 50 thousand devices, Wednesday with close to 45 thousand devices, Friday with just over 50 thousand devices, Saturday with close to 55 thousand devices, and Sunday with nearly 60 thousand devices.

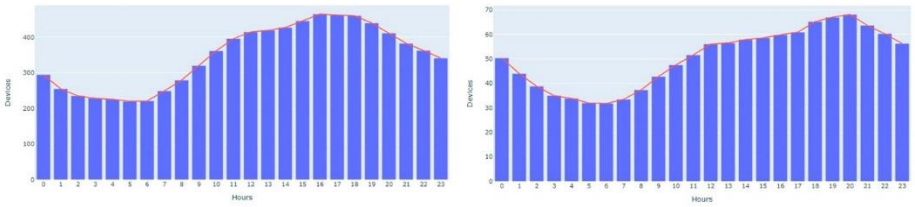

Fig. 6. Average thousand devices in December daily hours in Santa Maria Maior: non-roaming (left) and roaming (right).

In **Fig. 6**, we show two bar plots with the average number of non-roaming and roaming devices, in December's daily hours. For non-roaming data we verify that the minimum device peak is at 6 a.m. with slightly more than 200 thousand devices, and the maximum device peak is at 4 p.m. with around 450 thousand devices, showing that it took 10 hours to reach the maximum value. The common commute pattern considers the morning peak between 8 a.m. to 9 a.m. and at the afternoon peak between 5 p.m. or 6 p.m., which explains the observed behavior. For roaming data, we notice that the minimum device peak occurs at 6 a.m. with a little more than 30 thousand devices, and the maximum device peak occurs at 8 p.m. with nearly 70 devices, requiring 14 hours to achieve the maximum figure. Roaming mobile phone data shows a later appearance than non-roaming devices.

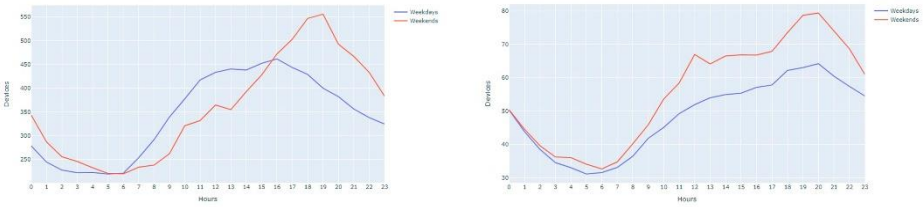

Fig. 7. Average thousand devices in December weekdays (blue line) and weekends (red line), per hours of the day, in Santa Maria Maior: non-roaming (left) and roaming (right).

In **Fig. 7**, we compare the average number of devices per hour of the day during the week (blue line) and on weekends (red line), using two-line plots for non-roaming (left) and roaming (right).

In the non-roaming graph, the contrast in the number of devices between the two categories outstood, with fewer devices on weekends. Following a pattern quite like that depicted in **Fig. 6**, the number of devices in Santa Maria Maior begins to increase at 6 a.m. in the morning for the working days of the week and weekends scenarios and continues to rise until 4pm for the working days the week and 5pm for the weekends, when the values begin to decrease. We confirmed that during the weekends, the devices only have higher values between 4 p.m. until 5 a.m. Moreover, Santa Maria Maior has more devices on working days during the day and on weekends during the night.

The roaming graph, with the two-line plots showed almost identical despite the difference during the day, as in the weekend showed highest values than the weekdays.

The weekday values began to increase at 5 a.m. and continued to rise until 8 p.m., at which point they begin to fall until 5 a.m. On the other hand, the weekdays started increasing at 6 a.m., peaking at 12 p.m., declining until 1 p.m., and rising again until 8 p.m., decreasing till 5 a.m. In this scenario, there are already more data on weekends than during the week.

Results and conclusions from graps' visualization are in line with expectations since Santa Maria Maior is one of most popular and touristics parishes in Lisbon.

Looking in more detail on POIs category analysis, we created visualizations using an ascending horizontal bar chart, with month by month average device figures, in each POI (see **Fig. 8**). The subway and train POIs were grouped together. As mentioned, if two POIs are paired with the same quadrant, they will have the same number of devices.

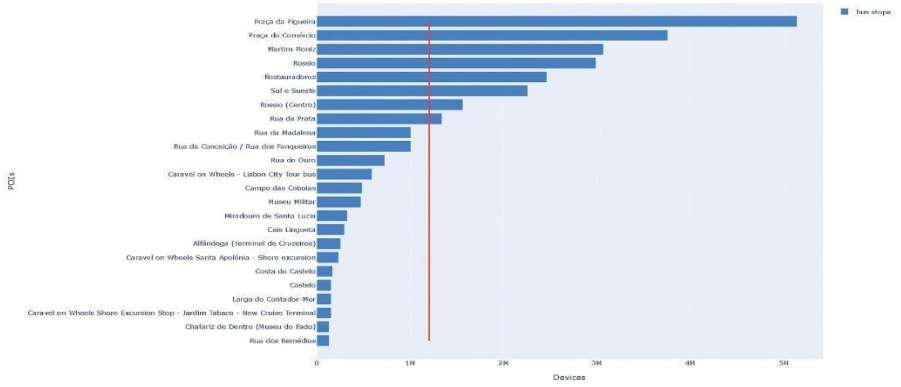

Fig. 8. Average bus stops POIs with devices in December, in Santa Maria Maior: roaming data.

In the non-roaming case Praça da Figueira had most of the devices, nearly 35 million, Martim Moniz, with over 25 million devices and Praça do Comércio, with nearly 25 million devices in December (the month with the highest number of devices). These areas of Santa Maria Maior correspond to high concentration areas of bus stops. The bus stops in the areas of Jardim do Tabaco, Rua dos Remédios and Chafariz de Dentro have the fewest number of devices, around 1 million each.

In the roaming graph (see **Fig. 8**), the top 3 areas were the same, with the order shift of 2nd and 3rd place: Praça da Figueira with more than 5 million devices, Praça do Comércio with just over 3.5 million, and Martim Moniz with just over 3 million. Again, the bus stop areas with the lowest number of devices were Jardim do Tabaco, Chafariz de Dentro, and Rua dos Remédios, with a total less than half a million.

When comparing the graphs between non-roaming and roaming, non-roaming shows an average of around 8 million, and roaming an average of approximately 1 million.

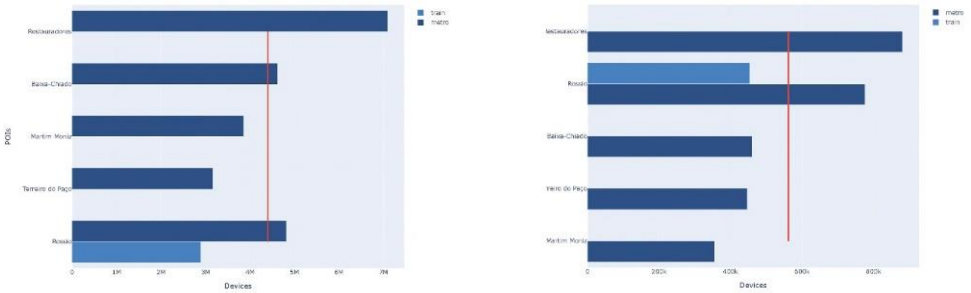

Fig. 9. Average train and metro stations POIs with devices in December, in Santa Maria Maior: non-roaming (left) and roaming (right).

In the non-roaming case the railway station with the most devices in December was Restauradores, together with the homonymous metro station, which has slightly more than 7 million devices, while the fewest devices were observed in the Rossio train station, with almost 2 million devices. These POIs contain an average devices of nearly 4.5 million.

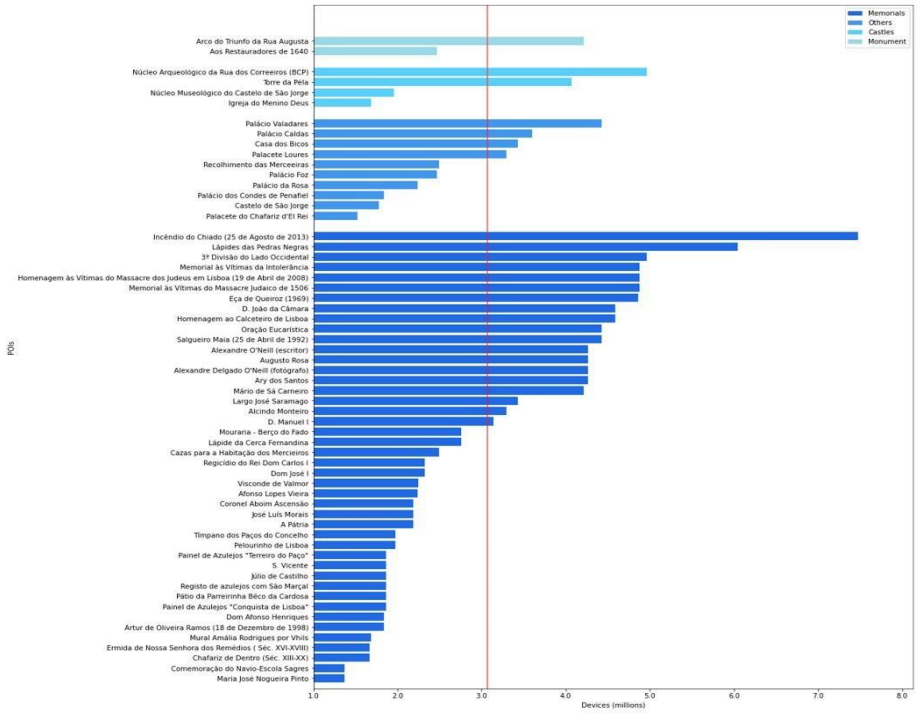

Fig. 10. Average historic POIs with devices in December, in Santa Maria Maior: non-roaming (left) and roaming (right).

In the roaming graph (see Fig. 10), between Rossio and Martim Moniz, Restauradores metro station had the highest number of devices with approximately 900,000 devices, followed by Rossio, with nearly 800,000 devices, Baixa-Chiado, Rossio, and Terreiro do Paco with values below 500,000 devices, and Martim Moniz with just under 400,000 devices. Comparing non-roaming and roaming, the non-roaming case has an average of around 4.5 million devices and the roaming case has an average of approximately 500,000.

Therefore, we modified our modelling approach and instead of having all our original groups — castle, memorial, monument, etc. — we created four groups: castle, memorial, monument + museum and others with the remaining groups (see **Fig. 10**).

In **Fig. 10**, we show that memorials had the highest number of historic POIs (44), followed by castles (10), others (4), and monuments (2), where other comprise churches, city walls, and archaeological sites. In both non-roaming and roaming cases, the memorial type not only has the largest number of POIs (more than half of POIs, but also showed the bigger presence of devices, with more than 2.5 million non-roaming, and 3 million roaming. The historical site with the most devices is the memorial of the Incêndio do Chiado (5.5 million non-roaming, and 7.5 million roaming), and the memorial of Maria José Nogueira Pinto and Comemoração do Navio-Escola Sagres are the sites with the fewest with less than 1.5 million devices (non-roaming and roaming).

Fig. 11. Cluster analysis with DBSCAN of historic, bus stops, metro, and train stations POIs with devices in December, in Santa Maria Maior: non-roaming (left) and roaming (right).

3.4.2 Cluster Analysis Model

In our modelling approach we aim also to identify correlations and structures in the data that would be difficult to find manually, but could also be useful in recognizing patterns and anticipating trends in our selected POIs. To this aim, we conducted cluster analyses with DBSCAN, for bus stops and historical points together, to achieve a sufficient data set size. The result is depicted in Fig. 11 where the color dots represent each of the eight found clusters for the month of December. Red dots represent POIs considered outliers, i.e., points which were discarded during clustering. The black cluster corresponds to Praça dos Restauradores, Praça Dom Pedro IV, and Praça da Figueira, in the Santa Justa area. The pink cluster regards Praça Martim Moniz, in the Mouraria neighborhood. Chiado and Sacramento correspond, respectively. to the dark

orange and deep pink clusters The longest cluster is colored purple and extends from Mártires to Castelo/Santiago. The yellow cluster is located in the Madalena neighborhood. The brown cluster corresponds to the neighborhood of Sé. The orange cluster includes the areas of Alfama and São Miguel. The blue cluster is located between São Miguel and Santo Estêvão. And close to the São Vicente parish, Santo Estêvão neighborhood, the green cluster evolves around the Lisbon Military Museum.

We highlight two of the found clusters: the purple and the blue. The purple cluster follows the route of the well-known electric tram 28E | Martim Moniz – Prazeres across the Baixa area, whereas the blue cluster is located in the Lisbon Cruise Terminal area bordering the green cluster in the Museu Militar. In Fig. 11 we can observe that, in these clusters, non-roaming mobility is higher.

3.5 Discussion

The literature review led to the collection of academic papers that determined the methodologies applied in this paper. Our study addressed the understanding of how travelers handling a mobile phone (nationals or roaming users), move in the city of Lisbon, particularly in the vicinity of POIs in the city, namely historic places and public transportation, with a special focus in the Santa Maria Maior parish. We analyzing georeferenced data collected by Vodafone during a five-month period, from September 2021 to January 2022 and open-source mapping data of the City [19] (OSMnx). We used statistical analysis and clustering analysis with DBSCAN, to investigate such travel phenomena. The development of various combinations between categories of POIs with DBSCAN [6]–[9] led us to conclude that the optimal result was eight clusters, of which two clusters, the purple cluster and the blue cluster, stood out due to their proximity to 28E tram route and Lisbon Cruise Terminal.

This research produced an innovative study taking the perspective of public transportation rather than shared transportation [16], with a focus case in the parish of Santa Maria Maior. This was chosen due to the large number of observed devices for non-roaming and roaming travelers, as well as the large number of POIs.

3.6 Research Limitations

We can highlight a few limitations, regarding the mobile phone data quality. The month of September data only begins on the 15th, making it an incomplete month. Additionally, the monthly datasets were encoded incorrectly, resulting in certain damaged values and improper formatting of the datetime and polygon objects. The data lacks also information on the nationalities of the roaming devices. The identification of the time they spend in a specific square polygon could result in an interesting analysis of trajectory patterns. The data corresponds to a pandemic-restricted season, which does not represent a usual mobility period. The inclusion of anonymous device identification could also allow a more extensive study to better understand the trajectories of travelers.

3.7 Future Work

Future work could cross-reference mobile phone data with public transportation cards (Viva/Navegante) data, in order to understand the entries and exits in transportation modalities. The availability the roaming nationality variable, would enable to comprehend distinct behavioral patterns from different nationalities. With a higher processing capacity, it would be possible to analyze all monthly data in a single dataset and generate more dynamic graphs, including the analysis of daily peaks, during the day, afternoon, or night, taking into account the full dataset.

Acknowledgements
This work is partially funded by national funds through FCT - Fundação para a Ciência e Tecnologia, I.P., under the project FCT UIDB/04466/2020.

References

[1] "Mobilidade na cidade de Lisboa com base em dados de telemóveis – LxDataLab." https://lisboainteligente.cm-lisboa.pt/lxdatalab/desafios/mobilidade-na-cidade-de-lisboa-com-base-em-dados-de-telemoveis/ (accessed Aug. 30, 2022).

[2] "LxDataLab - Lisboa Inteligente." https://lisboainteligente.cm-lisboa.pt/lxi-iniciativas/lxdatalab/ (accessed Sep. 03, 2022).

[3] "CRISP-DM - A Framework For Data Mining & Analysis." https://thinkinsights.net/data-literacy/crisp-dm/ (accessed Oct. 21, 2022).

[4] C. Schröer, F. Kruse, and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," *Procedia Comput Sci*, vol. 181, pp. 526–534, Jan. 2021, doi: 10.1016/J.PROCS.2021.01.199.

[5] M. J. Page *et al.*, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *The BMJ*, vol. 372, Mar. 2021, doi: 10.1136/BMJ.N71.

[6] M. Saliba, C. Abela, and C. Layfield, "Vehicular traffic flow intensity detection and prediction through mobile data usage," in *CEUR Workshop Proceedings*, 2018, vol. 2259, pp. 66–77.

[7] Irrevaldy and G. A. P. Saptawati, "Spatio-temporal mining to identify potential traff congestion based on transportation mode," in *Proceedings of 2017 International Conference on Data and Software Engineering, ICoDSE 2017*, 2018, vol. 2018-Janua, pp. 1–6. doi: 10.1109/ICODSE.2017.8285857.

[8] C. Li, J. Hu, Z. Dai, Z. Fan, and Z. Wu, "Understanding individual mobility pattern and portrait depiction based on mobile phone data," *ISPRS Int J Geoinf*, vol. 9, no. 11, 2020, doi: 10.3390/ijgi9110666.

[9] M. Li, B. Jin, H. Tang, and F. Zhang, "Clustering large-scale origin-destination pairs: A case study for public transit in Beijing," *Proceedings - 2018 IEEE SmartWorld, Ubiquitous Intelligence and Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People and Smart City Innovations,*

SmartWorld/UIC/ATC/ScalCom/CBDCo, pp. 705–712, 2018, doi: 10.1109/SmartWorld.2018.00137.

[10] S. Qin, J. Man, X. Wang, C. Li, H. Dong, and X. Ge, "Applying Big Data Analytics to Monitor Tourist Flow for the Scenic Area Operation Management," *Discrete Dyn Nat Soc*, vol. 2019, pp. 1–11, 2019, doi: 10.1155/2019/8239047.

[11] C. Balzotti, A. Bragagnini, M. Briani, and E. Cristiani, "Understanding Human Mobility Flows from Aggregated Mobile Phone Data∗," 2018, vol. 51, no. 9, pp. 25–30. doi: 10.1016/j.ifacol.2018.07.005.

[12] Y. Yuan and M. Raubal, "Extracting dynamic urban mobility patterns from mobile phone data," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012, vol. 7478 LNCS, pp. 354–367. doi: 10.1007/978-3-642-33024-7_26.

[13] P. Wang, J. Zhang, G. Liu, Y. Fuu, and C. Aggarwal, "Ensemble-spotting: Ranking urban vibrancy via POI embedding with multi-view spatial graphs," in *SIAM International Conference on Data Mining, SDM 2018*, 2018, pp. 351–359. doi: 10.1137/1.9781611975321.40.

[14] T. G. Martins, N. Lago, E. F. Z. Santana, A. Telea, F. Kon, and H. A. de Souza, "Using bundling to visualize multivariate urban mobility structure patterns in the São Paulo Metropolitan Area," *Journal of Internet Services and Applications*, vol. 12, no. 1, 2021, doi: 10.1186/s13174-021-00136-9.

[15] H. Senaratne *et al.*, "Urban Mobility Analysis with Mobile Network Data: A Visual Analytics Approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 5, pp. 1537–1546, 2018, doi: 10.1109/TITS.2017.2727281.

[16] T. Fontes, M. Arantes, P. v. Figueiredo, and P. Novais, "A Cluster-Based Approach Using Smartphone Data for Bike-Sharing Docking Stations Identification: Lisbon Case Study†," *Smart Cities*, vol. 5, no. 1, pp. 251–275, 2022, doi: 10.3390/smartcities5010016.

[17] S. A. Haidery, H. Ullah, N. Ullah Khan, K. Fatima, S. Shahla Rizvi, and S. J. Kwon, "Role of big data in the development of smart city by analyzing the density of residents in shanghai," *Electronics (Switzerland)*, vol. 9, no. 5, 2020, doi: 10.3390/electronics9050837.

[18] "Diário da República, 1.ª série — N.º 216 — 8 de novembro de 2012 ." 2012. Accessed: Sep. 09, 2022. [Online]. Available: https://files.dre.pt/1s/2012/11/21600/0645406460.pdf

[19] "OSMnx 1.2.2 — OSMnx 1.2.2 documentation." https://osmnx.readthedocs.io/en/stable/ (accessed Sep. 09, 2022).