INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Mobility in Lisbon based on smartphone data

Daniel Romão Leal

Master in **Integrated Business Intelligence Systems**

Supervisor:
PhD José Miguel de Oliveira Monteiro Sales Dias, Associated Professor with Habilitation
ISCTE-IUL

Co-supervisor:
MSc Lídia Vitória Pires de Albuquerque, Researcher
NOVA IMS

October, 2022

iscte

TECNOLOGIAS
E ARQUITETURA

Department of Information Science and Technology

Mobility in Lisbon based on smartphone data

Daniel Romão Leal

Master in **Integrated Business Intelligence Systems**

Supervisor:
PhD José Miguel de Oliveira Monteiro Sales Dias, Associated Professor with Habilitation
ISCTE-IUL

Co-supervisor:
MSc Lídia Vitória Pires de Albuquerque, Researcher
NOVA IMS

October, 2022

“Face the demands of life voluntarily. Respond to a challenge, instead of bracing for catastrophe.”

Jordan Peterson

Acknowledgments

Genuinely, I appreciate the time and effort invested by Professor Miguel Sales Dias and Vitória Albuquerque, my thesis mentors and supervisors, without whom this thesis would have not been possible. We were able to communicate well, and they were always there when I needed them to answer questions or give guidance. They set a standard for excellence in their field and mentored those around them. As a result of my research, I am better informed and have developed a more efficient work routine. I also want to thank Professor João Ferreira for introducing and inspiring me to work on this topic and for his support. I would like to thank the Câmara Municipal de Lisboa, in collaboration with LxDataLab and Vodafone, for supplying me with the data used in this thesis.

Resumo

Este estudo abrange cinco meses (setembro, outubro, novembro, dezembro de 2021 e janeiro de 2022) de dados georreferenciados do serviço da operadora móvel Vodafone, fornecido pela Câmara Municipal de Lisboa (CML). A motivação da tese considera o facto de o estudo da mobilidade urbana com dados de telemóveis ser um tópico relativamente inexplorado. Este estudo centrou-se na cidade de Lisboa, com um caso de estudo na freguesia de Santa Maria Maior com o objetivo de compreender os padrões de mobilidade urbana dos utilizadores da rede móvel. O número de dispositivos de não-roaming e roaming no caso de estudo está relacionado com o tema das 'vibrant neighborhoods' e turismo, caracterizado por pontos de interesse históricos e de transportes. Utilizámos uma abordagem de 'data mining' para analisar as tendências de mobilidade, adotando uma metodologia CRISP-DM, para realizar análise estatística, visualização e agrupamentos (DBSCAN). Os resultados mostraram nove agrupamentos em Santa Maria Maior, dos quais dois agrupamentos de destaque, um ao longo do elétrico 28-E e outro à volta do Terminal de Cruzeiros de Lisboa. Em primeiro lugar, analisámos estes dois agrupamentos e realizámos análises de previsão, resultando numa tendência decrescente, como consequência das restrições da pandemia nos meses de dezembro e janeiro. Esta tese contribui consideravelmente para a transformação digital de Lisboa numa cidade inteligente, ao compreender os padrões de mobilidade urbana com dados dos utilizadores da rede móvel em não-roaming e roaming.

Palavras-chave: operadora móvel, padrões de mobilidade, visualização, vibrant neighborhoods, pontos de interesse, DBSCAN

Abstract

This research covers five months (September, October, November, December 2021, and January 2022) of georeferenced data of the Vodafone mobile phone service, provided by the municipality of Lisbon (CML). The motivation of this research regards the fact that the urban mobility study with mobile phone data is a relatively unexplored topic. This study focused on the city of Lisbon, with a case study conducted in the parish of Santa Maria Maior with the aim to understand the urban mobility patterns of mobile phone users. The number of roaming and non-roaming devices in the case study is related to the subject of a vibrant neighborhood and tourism, characterized by transportation and historical points of interest. We used a data mining approach to analyze mobility trends, adopting a CRISP-DM methodology, to perform statistical analysis, visualization, and clustering (DBSCAN) methods. Results showed eight clusters in Santa Maria Maior, with outstanding clusters along 28-E electric tram and Lisbon Cruise Terminal. Foremost, we looked at these two clusters and performed a forecast model with Prophet, resulting in downward trend, influenced by the pandemic restrictions in December and January data. This thesis contributes considerably to the digital transformation of Lisbon into a smart city by understanding urban mobility patterns with smartphone data of no roaming and roaming users.

Keywords: smartphone data, mobility patterns, visualisation, vibrant neighbourhoods, point of interest, DBSCAN

Index

Acknowledgments iii

Resumo v

Abstract vii

Index ix

Tables index xi

List of abbreviations xiv

1. Introduction 1

1.1. Topic context and motivation 1

1.2. Research questions and objectives 2

1.3. Methodological approach 2

1.4. Dissertation structure and organization 3

2. Literature review 5

2.1. Context 5

2.2. Literature review methodology 5

2.2.1. Keyword and research query 5

2.2.2. PRISMA results 6

2.3. VOSviewer network analysis and visualization 9

2.3.1. Title and abstract analysis 9

2.3.2. Author co-authorship analysis 11

3. Smartphone data mobility patterns analytics 12

3.1. Business understanding 12

3.2. Data understanding 12

3.3. Data preparation 14

3.4. Data modeling 21

3.4.1. Smartphone data analysis and visualization 22

3.4.2. Smartphone data and POIs analysis and visualization 24

3.4.3. Smartphone data and POIs clustering 31

3.4.4. Forecasting with the Prophet algorithm 35

3.5. Evaluation 40

3.6. Deployment 42

4. Conclusions 43

4.1. Discussion 43

4.2. Research limitations 45
4.3. Future work 45
References 47
Appendix 51

Tables index

Table 1 - Literature review application and method 8
Table 2 – Monthly datasets 13
Table 3 – Geoinformation dataset........ 14
Table 4 – Table of resulted monthly datasets 17
Table 5 – Method assessment questionnaire*........ 41

Figures INDEX

Figure 1 - CRISP-DM diagram 3
Figure 2 – PRISMA flow diagram 7
Figure 3 - Title and abstract network visualization 10
Figure 4 - Title and abstract network overlay visualization 10
Figure 5 - Author co-authorship network overlay visualization 11
Figure 6 – Boxplots of numeric variables in non-roaming (left) and roaming (right) December dataset 15
Figure 7 – Boxplots of numeric variables in September dataset 16
Figure 8 – Devices (millions) per parish in December non-roaming (left) and roaming (right) dataset18
Figure 9 – Historic POIs per parish 19
Figure 10 – Bus stop POIs per parish 20
Figure 11 – Metro station POIs per parish 20
Figure 12 – Train station POIs per parish 21
Figure 13 – Average devices (thousands) over December’s non-roaming (left) and roaming (right) weekdays in Santa Maria Maior 22
Figure 14 – Average devices (thousands) over December’s non-roaming (left) and roaming (right) daily hours 22
Figure 15 – Average devices (thousands) for weekdays (blue line) and weekends (red line) over December’s non-roaming (left) and roaming (right) daily hours 23
Figure 16 – Historic POIs influence area (red dots) in Santa Maria Maior’s smartphone data quadrants in December non-roaming (left) and roaming (right) 25
Figure 17 – Bus stop POIs influence area in Santa Maria Maior’s quadrants in December non-roaming (left) and roaming (right) 25
Figure 18 – Metro and train POIs influence area in Santa Maria Maior’s quadrants in December non-roaming (left) and roaming (right) 25
Figure 19 – POIs category histogram 26
Figure 20 – POIs location distribution in Santa Maria Maior 27
Figure 21 – Santa Maria Maior historic POIs with non-roaming devices’ average (red line) in December roaming 28
Figure 22 – Santa Maria Maior bus stops POIs with devices’ average (red line) in December roaming 29
Figure 23 – Santa Maria Maior train and metro stations POIs with devices’ average (red line) in December non-roaming (left) and roaming (right) 30
Figure 24 – Diagram of DBSCAN point types 31
Figure 25 – The Elbow method for historic, bus stops, metro, and train stations POIs 32
Figure 26 – DBSCAN historic POIs influence area (colored dots) in Santa Maria Maior’s smartphone December’s non-roaming (left) and roaming (right) data quadrants (millions) 33
Figure 27 – DBSCAN bus stops POIs influence area (colored dots) in Santa Maria Maior’s smartphone December’s non-roaming (left) and roaming (right) data quadrants (millions) 34
Figure 28 – DBSCAN historic and bus stops POIs influence area (colored dots) in Santa Maria Maior’s smartphone December’s non-roaming (left) and roaming (right) data quadrants (millions) 34

Figure 29 – DBSCAN historic, bus stops, metro, and train stations POIs influence area (colored dots) in Santa Maria Maior’s smartphone December’s non-roaming (left) and roaming (right) data quadrants (millions) 34
Figure 30 – Time series of blue cluster (first row) and purple cluster (second row) non-roaming (left) and roaming (right) during all months’ dataset 37
Figure 31 – Blue cluster (first row) and purple cluster (second row) for non-roaming (left) and roaming (right) monthly with February prediction (blue shade), weekly and hourly graphs during all months’ dataset 39
Figure 32 – Prophet plot for blue cluster (first row) and purple cluster (second row) for non-roaming (left) and roaming (right) during all months’ dataset with time series (black dots), the forecasted values (blue line), the trend values (red line), the projected uncertainty intervals (blue shaded regions) and the potential changepoints (vertical red lines) 40

List of abbreviations

CML - Câmara Municipal de Lisboa

CRISP-DM - Cross-Industry Standard Process for Data Mining

DBSCAN – Density-Based Spatial Clustering of Applications with Noise

DDos – Distributed Denial-of-Service

DTW – Dynamic Time Warping

GPS – Global Position System

GSM – Global System for Mobile

HTML – HyperText Markup Language

Iscte – Instituto Superior de Ciências Sociais do Trabalho e da Empresa

IoT – Internet of Things

POI – Point Of Interest

PRISMA - Preferred Reporting Items for Systematic Reviews and Meta-Analyses

RAM – Random Access Memory

RQ - Research Question

SDGs - Sustainable Development Goals

SIM – Subscriber Identity Model

UMTS - Universal Mobile Telecommunications System

UN-GA - United Nations General Assembly

WKT - Well-known Text

1. Introduction

1.1. Topic context and motivation

This thesis intends to address two of the seventeen Sustainable Development Goals (SDGs) [1] established by the United Nations General Assembly (UN-GA) [2] in 2015, with a target date of 2030. The Sustainable Development Goals we tackle target within the following themes: Innovation and Infrastructure, and Sustainable Cities and Communities. The two SDGs provide an understanding to enable the exchange of information on the role that smart cities may play in enhancing global sustainability.

This study aims to improve the planning and management of cities to improve the quality of life of citizens in a sustainable manner by making the best use of data, technology, and innovative technical resources available.

The analysis of Internet of Things (IoT) data shows we alleviate urbanization’s pressures by providing a new experience for city residents and making day-to-day life more comfortable and secure. In smart cities, the IoT refers to the use of smart technology and linked devices for real-time data collection. Rising urbanization, increased demand for efficient infrastructure in metropolitan areas, increased demand for energy-efficient resources, traffic management, waste management, public safety, and security are all development factors for the total market. Connected internet technologies can be used to alleviate problems, improve the quality of life of residents, and minimize resource consumption in smart cities.

It has become increasingly crucial to determine the location of mobile users in Global System for Mobile (GSM) networks. The location of a cell phone can be determined using the network architecture of the 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’s 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. 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 a familiar part of our daily routines over the years, as is the struggle for research to help reduce this. As a result, it is becoming increasingly vital to revolutionize traffic management in urban areas using data in a variety of methods to help cities get a clearer view of what is happening and adapt to provide new mobility solutions.

With the Vodafone data provided by Câmara Municipal de Lisboa (CML), there is an opportunity and interest to study this data in the scope of people's mobility in the city of Lisbon, especially how they travel over time. Considering this data, our aim in this research is to understand mobility patterns in Lisbon, by performing analysis and visualization of the Vodafone users' data.

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

1.2. Research questions and objectives

This research theme was proposed by Instituto Superior de Ciências do Trabalho e da Empresa (Iscte) in partnership with CML's Center for Management and Urban Intelligence [3] by the LxDataLab [4] coordinated by the National Scientific and Technological System (SCTN) [5] in partnership with Vodafone Portugal.

The motivation for this study is the few existing studies on the subject as well as for this case study.

Our main 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 in the city, namely historic places and public transportation?"

This led us to our research objective that is, in short, to understand the mobility patterns in Lisbon through mobile phone data, in the vicinity of the mentioned points of interest. We proposed to perform analysis and visualization of mobile Vodafone data to identify mobility patterns in Lisbon, using data mining and visualization. In our data mining approach, adopting the CRISP-DM methodology [6], [7], we will use statistical analysis and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) method.

This led us to very interesting results that allowed us to move a step forward and propose a sub-question: "How can we forecast mobility patterns of smartphone users in Santa Maria Maior, Lisbon?"

From our DBSCAN analysis, we built a forecasting algorithm with Prophet implementation, with the objective to understand future patterns in mobility of mobile phone users (non-roaming and roaming).

1.3. Methodological approach

In this dissertation, we applied the Cross Industry Standard Process for Data Mining (CRISP-DM)[6], [7] as shown in Figure 1. This process consists of six sequential phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment.

The “Business Understanding” phase is about getting a general overview of the project’s goals and requirements. A solid understanding of the customer’s business goals is necessary before initiating the process of determining the business goals and then determine what constitutes a successful business. Next step is “Data Understanding”, this is where we will analyze, identify, and collect the data to achieve the project aim. “Data Preparation” is the phase where the final data set(s) is ready for modelling: determine which data sets and documents will be used, cleaning the data, data construction, data integration and, finally, data format. In the fourth phase, “Data Modelling”, consists in the development and evaluation of a variety of models using a different modelling strategy. In the “Evaluation” phase, we evaluate, review, and determine the subsequent steps based on outcomes. In the last step, “Deployment” we make a final report on our findings, with a summary of the project and plan thoroughly for monitoring and maintenance for future integrations.

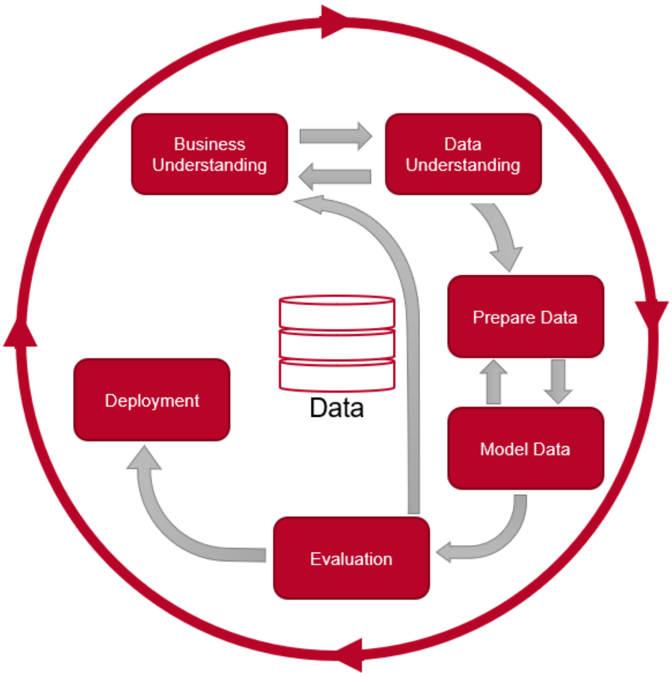


Figure 1 - CRISP-DM diagram

1.4. Dissertation structure and organization

This dissertation is organized into four sections. In section 1, we introduce the topic context and motivation, research questions and objective, methodology, and structure. Section 2 presents the bibliometric analysis using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [8] to find the latest state-of-the-art methodologies applied to mobility behavioral patterns with GSM data. Then, we apply the VOSviewer [9] tool to visualize scientific landscapes in our literature dataset, as well as title and abstract words occurrence. Section3, we apply the CRISP-DM methodology to our Vodafone dataset with a data science approach to perform data mining, statistical analysis, and

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering. In section 4, we present our conclusions discussion, limitations, and future work.

2. Literature review

2.1. Context

Research domains such as geographic information science, transportation, and physics have turned human movement patterns modelling into an essential study subject.

Yuan [10] analyzed data from individual traveling and tracked them using GPS, GSM, and other locators, emphasizing physical travel of people, discussing human mobility.

Understanding better the behavior of urban mobility patterns enables us to conduct an analysis and aid a better comprehension of the dynamic nature of various metropolitan areas, as well as the revision of environmental and transportation policy.

In our case study, the data was provided by only one operator, Vodafone, as such it does not reflect all the existing individual mobility information, as not all individuals belong to the same operator, preventing us from achieving the trend of a more effective and real data analysis.

The objective of this thesis is to extract mobility-related behavioral patterns and analyze their spatiotemporal features by combining calendar events and points of interest (POI).

2.2. Literature review methodology

PRISMA [8] was applied with the purpose of identifying, evaluating, and critically appraising research to provide an answer to a well-formulated query. This methodology is a minimal set of elements for systematic reviews and meta-analyses that is scientific proof. It is composed of a 27 items checklist and a four-phase flow diagram.

We conducted our search in April 2022 and restricted our examination of the scientific literature to articles published during the last five years, from 2017 to 2021, in English.

2.2.1. Keyword and research query

Researching technical phrases used by the scientific community and using trial and error were the components in this stage that resulted in the creation of the most suitable query for the thesis theme.

In the search, keywords such as Data Mining and Machine Learning, Traffic Congestion, Smartphone, and Data; and articles published between 2017 and 2021 were the ones that, when used featured with the appropriate operators, returned the most useful results.

Following our investigation, we came to the following queries for the two platforms, as their syntaxes varied:

Scopus Query:

(("Data Mining" OR "Machine Learning") AND ("Traffic Congestion" OR "Road") AND (“SMARTPHONE” OR "MOBILE") AND "DATA"))

Web of Science Query:

((AB= "Data Mining" OR AB="Machine Learning") AND (AB= "Traffic Congestion" OR AB="Road") AND (AB="SMARTPHONE" OR AB = ”MOBILE”) AND AB="DATA")

2.2.2. PRISMA results

We used the PRISMA Flow Diagram considering four steps: identification, screening, eligibility and included.

In the first PRISMA step, identification, we examined two database platforms for academic paper abstracts and citations: Scopus and Web of Science. To locate the desired articles, platform-specific queries were utilized 375 papers were obtained from the Scopus platform and 153 articles from Web of Science. The supervisors suggested 6 articles to be added to the total collected. Mendeley [11] software was used to assist in the collection of these articles. It is used to manage, distribute, and produce bibliographic references for academic papers. After all articles were added to Mendeley, 51 duplicates were eliminated. This resulted in a total of 355 papers.

We screened the papers’ titles and abstracts related to our research issue during the screening step. We consider articles to be rejected based on the title and abstract screening. Afterward, on the “Eligibility” step, a review of the content of these articles was conducted to determine whether they were eligible to be included in the “Included” step of this systemic review. In this situation, a total of 36 articles were taken into consideration. In this step, out-of-scope items were identified and grouped into agriculture, computer vision (3D approach, autonomous vehicles, AI), data management network (Distributed Denial-of-Service (DDoS) attacks, 5G network) and analysis of data sources, mobility management (mobility prediction, parking), safety (vehicle safety, accidents, wheelchair, driver behavior) and social (emotion, pollution, tourists).

Once we finished all the steps of the process, shown in our PRISMA flow diagram (Figure 2), we reached the result of 12 articles reviewed and considered in our research.

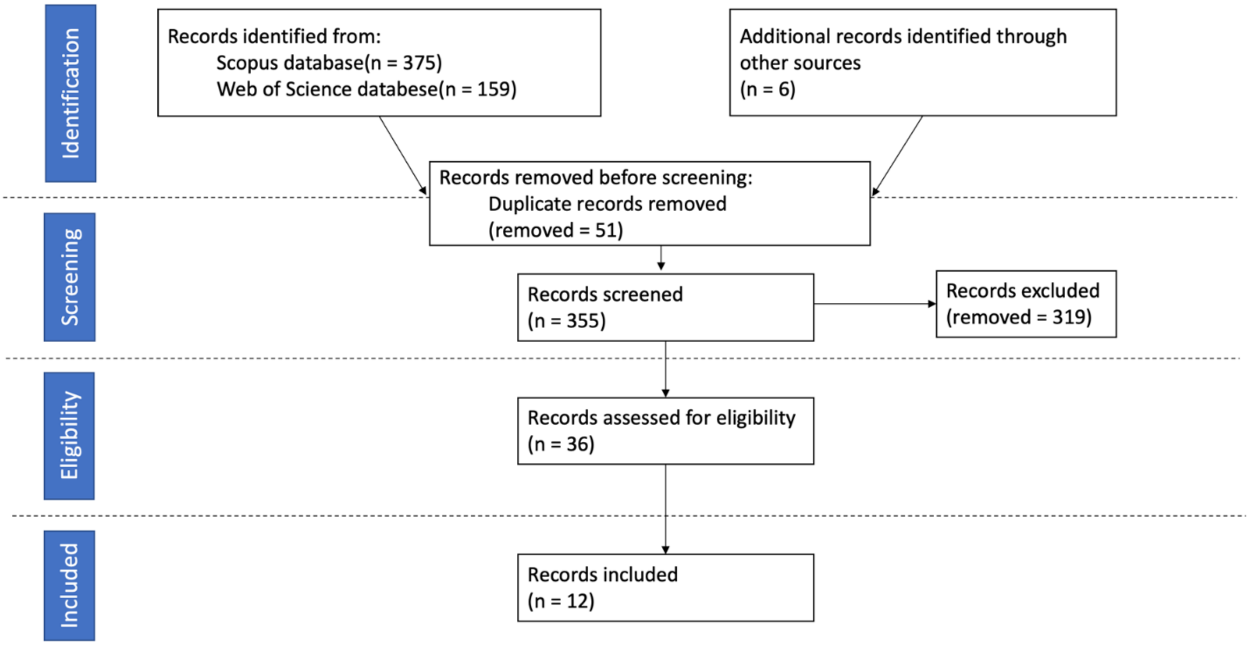


Figure 2 – PRISMA flow diagram

Based on the articles resulting from our PRISMA flow diagram we analyzed the methods and applications (Table 1). We observed a trend in the application of DBSCAN method [12]–[15]. Other methods used include visualization and analysis of mobility patterns with and without point of interest (POI) [10], [16]–[20], k-means clustering algorithm [21], Dynamic Time Warping (DTW) [10], and analytic methods such as Point Density and Kernel Density Estimation [22].

We identified the use of the DBSCAN method in "Vehicular traffic flow intensity detection and prediction through mobile data usage" [12] 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" [21] for the identification of soft mobility hotspots at specific bike share docking stations using k-means clustering algorithms; "Spatio-Temporal Mining To Identify Potential Traffic Congestion Based On Transportation Mode" [13] for the Identification of potential traffic congestion using DBSCAN clustering algorithm; “Understanding individual mobility pattern and portrait depiction based on mobile phone data” [14] with application for individual mobility pattern analysis and portrait depiction in various China cities; and "Clustering Large-Scale Origin-Destination Pairs: A Case Study for Public Transit in Beijing" with application to determining the mobility patterns of bus passengers in Beijing [15].

We also identified visualizations and people analysis in the articles "Applying Big Data Analytics to Monitor Tourist Flow for the Scenic Area Operation Management" [16] in which it is applied to the identification of tourist movement in Beijing; "Understanding Human Mobility Flows from Aggregated

Mobile Phone Data" [17] in which it is applied to the identification of population behavior in Milan; "Extracting Dynamic Urban Mobility Patterns Phone Data" [10] in which it is applied to the identification of urban mobility patterns. The research "Ensemble-spotting: Ranking urban vibrancy via POI embedding with multi-view spatial graphs" [18] revealed the application of the study of mobility patterns with POIs to the discovery of 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"[19] identified spatial grouping and some visualization using the application of bundling approach to support multi-attribute trail datasets in the São Paulo metropolitan area.

With the application of identifying urban mobility patterns in the city of Shanghai, an article of analytical approaches such as Point Density and Kernel Density Estimation titled "Role of big data in development of smart city by studying the density of citizens in Shanghai" [22] is considered.

Table 1 - Literature review application and method

Title	Authors	Application	Method
Vehicular traffic flow intensity detection and prediction through mobile data usage [12]	Saliba, M. Abela, C. Layfield, C.	An artificial neural network was trained with grid nodes' traffic levels in a timeseries to forecast node traffic	DBSCAN clustering
A cluster-Based Approach Using Smartphone Data for Bike-Sharing Docking Stations Identification: Lisbon Case Study [21]	Tiago Fontes, Miguel Arantes, Paulo V. Figueired, Paulo Novais	Identify soft mobility hotspots in specific bike-sharing docking stations using clustering algorithms	K-means
Spatio-Temporal Mining To Identify Potential Traffic Congestion Based On Transportation Mode [13]	Irrevaldy Saptawati, Gusti Ayu Putri	Identification of potential traffic congestion using clustering algorithm	DBSCAN clustering
Applying Big Data Analytics to Monitor Tourist Flow for the Scenic Area Operation Management [16]	Siyang Qin, Jie Man, Xuzhao Wang, Can Li, Honghui Dong, Xinquan Ge	Identification of tourist movement in Beijing	Visualization and analysis of tourists throughout time and spatial distribution in certain zones
Understanding Human Mobility Flows from Aggregated Mobile Phone Data [17]	Caterina Balzotti, Andrea Bragagnini, Maya Briani, Emiliano Cristiani	Identification of population behavior in Milan	Visualization and analysis of travel flows and patterns of people
Extracting dynamic urban mobility patterns from mobile phone data [10]	Yihong Yuan, Martin Raubal	Identifying urban mobility patterns in a city in China.	Analyzing human trajectories and motion patterns based on dynamic

			time warping (DTW) algorithm
Ensemble-spotting: Ranking urban vibrancy via POI embedding with multi-view spatial graphs [18]	Wang, P. Zhang, J. Liu, G. Fuu, Y. Aggarwal, C.	Identifying relation in vibrant communities with geographical items	Analyzing mobility patterns with POI's
Role of big data in the development of smart city by analyzing the density of residents in shanghai [22]	Haidery, S.A. Ullah, H. Ullah Khan, N. Fatima, K. Shahla Rizvi, S. Kwon, S.J.	Identifying urban mobility patterns in Shangai.	Analytic methods as Point Density and Kernel Density Estimation
Understanding individual mobility pattern and portrait depiction based on mobile phone data [14]	Li, C. Hu, J. Dai, Z. Fan, Z. Wu, Z.	Individual mobility pattern analysis and portrait depiction in various China cities	DBSCAN clustering
Using bundling to visualize multivariate urban mobility structure patterns in the São Paulo Metropolitan Area [19]	Martins, T.G. Lago, N. Santana, E.F.Z. Telea, A. Kon, F. de Souza, H.A.	Bundling technique to support multi-attribute trail datasets in São Paulo metropolitan area	Spatial clustering and visualization of urban mobility
Urban Mobility Analysis with Mobile Network Data: A Visual Analytics Approach [20]	Senaratne, H. Mueller, M. Behrisch, M. Lalanne, F. Bustos-Jimenez, J. Schneidewind, J. Keim, D. Schreck, T.	Identify human behavioral patterns in Santiago	Data visualization and data analysis algorithms
Clustering large-scale origin-destination pairs: A case study for public transit in Beijing [15]	Li, M, Jin, B. Tang, H. Zhang, F.	Identify the mobility patterns of bus passenger in Beijing	DBSCAN clustering

2.3. VOSviewer network analysis and visualization

2.3.1. Title and abstract analysis

In Figure 3 and Figure 4 we considered two VOSviewer network analysis visualization processes. The fields in each phase were selected with the intention of generating the optimal linking graph between terms included in the selected papers. Our chosen counting method was full counting.

In Figure 3 and Figure 4 we see that the term "data" is the one that stands out as the most significant one. It includes links to all the terms of the cluster.

The term analysis of the title and abstract produced 19 selected items with 4 clusters and 97 linkages for a total link weight of 620 terms.

In both Figure 3 and Figure 4 the term "point" is linked to "cluster" and "vibrant community," which leads us to the POI methodology or even "public transportation," which is related to "behavior trajectory”. On the other hand, I have a link between the terms "data" with "model", "time", "intensity", and "density", which is linked to the term "cluster".

On the temporal level of terms used between 2018 and 2020, it is possible to deduce that the words used at the beginning of 2018 were "vibrant community", "public transit", "cluster", "mobility pattern", "model", "urban area", and "time"; and the words used in the second half of that year were "trajectory", "point", "data" and "location". Already in 2019, the terms "density", "city", "smart city", and "beijing" were collected.

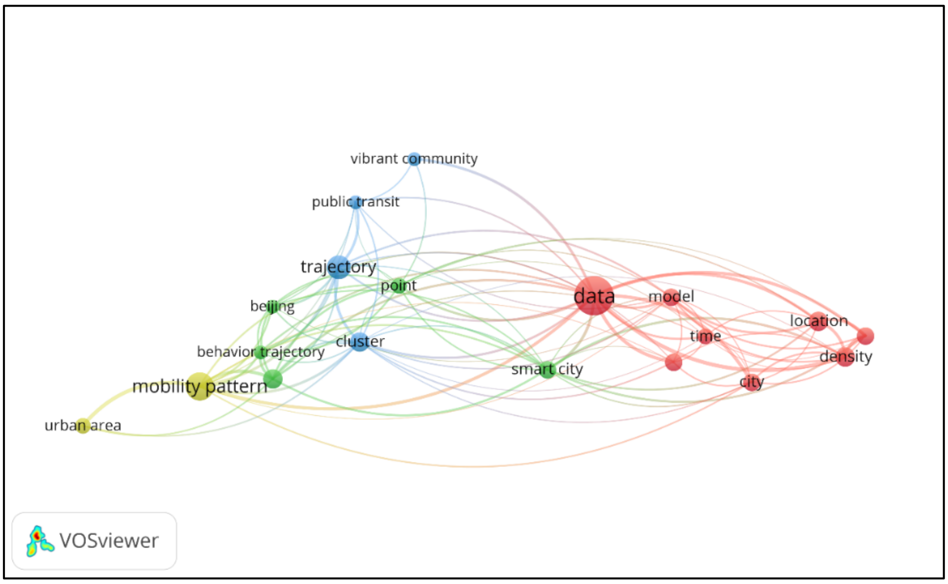


Figure 3 - Title and abstract network visualization

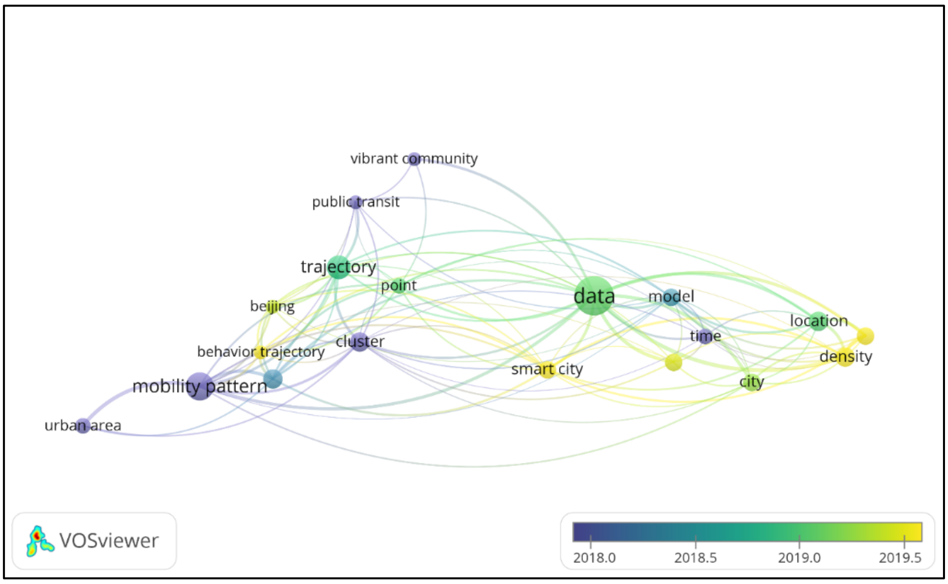


Figure 4 - Title and abstract network overlay visualization

2.3.2. Author co-authorship analysis

In Figure 5 demonstrates that most clusters are related to individual papers and the academic community does not collaborate with one another.

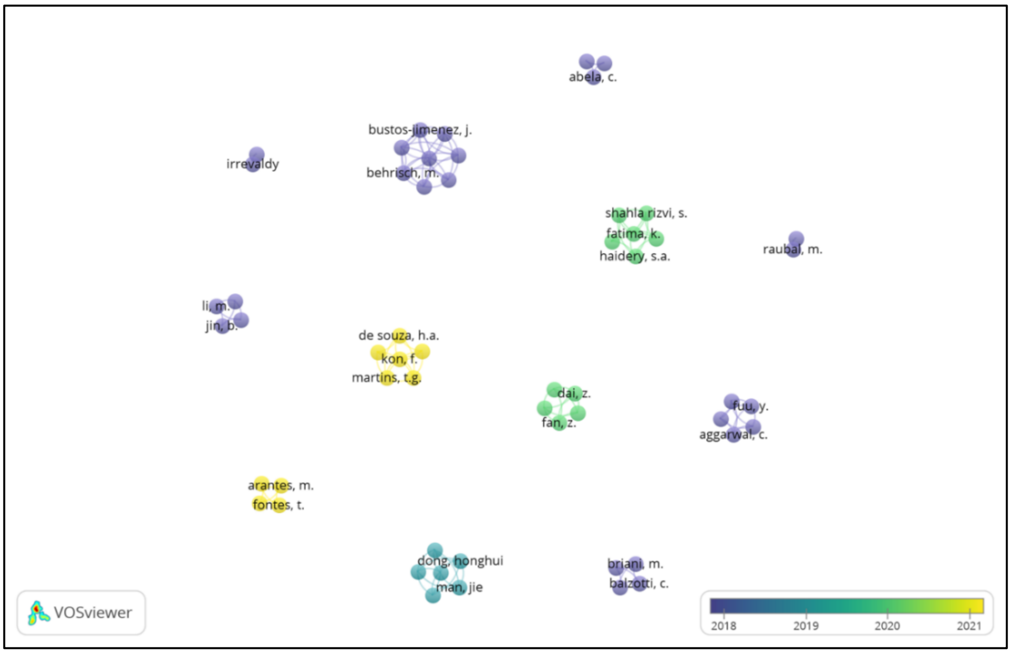


Figure 5 - Author co-authorship network overlay visualization

3. Smartphone data mobility patterns analytics

3.1. Business understanding

LxDataLab [4] 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 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 7 theme on “Mobility in the city of Lisbon based on mobile phone data” [23] and 71 theme on “Mobility evolution towards uplift of pandemic restrictions” [24] . This challenge in the mobility domain, in collaboration with a mobile service operator (Vodafone), aims to understand how people, handling a mobile phone, move in the city.

This thesis 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 questions. In short, 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 five months – 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 provided the number of devices presented 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 had seventeen million (17,233,318) records in September, thirty-two million (32,627,308) in October, twenty-one million (21,619,292) records in November, thirty-three million (33,121,657) records in December, and thirty-three million (33,344,624) records in January. This resulted in 137 million records for the five months.

The analysis of whether a certain device was in the grid, how long it was there or its route, was not possible due to the lack of information regarding the device identification. Since the goal of our study was to detect stationary devices in a particular grid, we chose to consider the values in columns C3 and C4, which reflected the devices that remained in each grid longer than 5 minutes for non-roaming and roaming, respectively.

The size of the datasets corresponding to each month added to 38GB, which was a significant computational limitation for the research, as our random-access memory (RAM) only has 32GB, making it impossible to cache the data records as well as the processing of our workstation for the analysis. As a result, the study of each five months was performed separately, in a distinct dataset for non-roaming and roaming, resulting in a total of 10 datasets for analysis.

An additional dataset with geoinformation, known as Vodafone grid, was provided by CML, and combined with the monthly datasets. This dataset complements the monthly ones 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 [25]. As such, we used the updated parishes information, set up after 2012.

Based in provided metadata by CML we constructed Table 2 Table 3.2 with information of columns from our monthly datasets and Table 3 with information of columns from Vodafone grid dataset.

Table 2 – Monthly datasets

Column	Description
Grid_ID	Row identifier
Datetime	Object module that supplies classes for manipulating dates and times
extract_year_2	Number of the year
extract_month_3	Number of the month
extract_day_4	Number of the days
C1	No. of distinct terminals in the grid
C2	No. of distinct terminals, roaming, in the grid
C3	No. of distinct terminals remaining in the grid
C4	No. of distinct terminals remaining in the grid, roaming
C5	No. of distinct terminal entries in the grid
C6	No. of outputs from different terminals in the grid
C7	No. of entries of distinct terminals roaming in the grid
C8	No. of exits from different terminals, roaming, in the grid cell
C9	No. of distinct terminals with active data connection, in the grid cell
C10	No. of distinct terminals with active data connection, roaming, in the grid cell
C11	No. of voice calls originating from the grid
C12	No. of entries into Lisbon along the 11 main roads
C13	No. of exits from Lisbon along the 11 main roads".
D1	Top 10 home countries of terminal equipment roaming
E1	No. of voice calls terminated in the grid
E2	Average downstream rhythm of the grid, in grid

E3	Average upstream rhythm of the grid
E4	Peak downstream rhythm of the grid
E5	Peak upstream rhythm of the grid
E6	Top 10 Applications (Text with names separated by;)
E7	Duration of the minimum stay within the grid
E8	Average length of stay in the grid
E9	Maximum duration of stay within the grid
E10	No. of devices performing grid connection sharing

Table 3 – Geoinformation dataset

Column	Description
grelha_id	Grid identifier
dicofre	-
entity_id	-
entity_type	-
freguesia	Parish name before new policies
freguesias	Parish name after new policies
grelha_x	-
grelha_y	-
latitude	Latitude value
longitude	Longitude value
nome	Location name
objectid	-
position	Geometric polygon in GeometryCollection type
wkt	Geometric multipolygon

3.3. Data preparation

After importing each of the 5 months datasets, we observed that all months except September had issues on csv file reading due to its encoding. It was easily fixed by adding the encoding attribute 'latin1' in the csv file reading procedure. The values in the datetime column were encoded in the datasets and had encoding issues (all datasets except September dataset) while reading, resolved after changing the datetime column as a basic object type to datetime64 object.

The columns that were not relevant, such as C1, C2, C5, C6, C7, C8, C9, C10, C11, D1, E1, E2, E3, E4, E5, E6, E7, E8, E9 and E10, to this analysis were removed due to the reasons stated in 3.2. Following this column cleaning, the initial dataset is formed with the Grid_ID, Datetime, extract_year_2, extract_month_3, extract_day_4, and column C3 or C4 (for not roaming or roaming). Our Grid_ID represented the identification of our data frame, the extract_year_2 the year, the extract_month_3 the month, the extract_day_4 the day, and the column C3 or C4 the number of devices in a specific square.

We included in our dataset three ordinal qualitative variables - extract_year_2, extract_month_3, and extract_day_4 - and three continuous variables – Grid_ID, Datetime, and C3 or C4.

The datasets (September, November, December, and January) contained 44 nulls in column C3/C4, 30 nulls in column extract_year_2, and 43 nulls in column extract_day_4. We removed their entries as they represented a small percentage of nulls in the datasets.

While working on the datetime object's encoding problem, we found issues in the October dataset with null values in the columns extract day 4, C3 (non-roaming cases) or C4 (roaming cases), Datetime, a day of the month value, in the extract day 4 column, of 22,820, and in the December dataset the column correspondent to the days, extract day 4, with a value of 0. Because of the size of the datasets in comparison to the number of records to be examined for deletion, the decision was to remove these records. So, the values in Figure 6 resulted from a cleaning process.

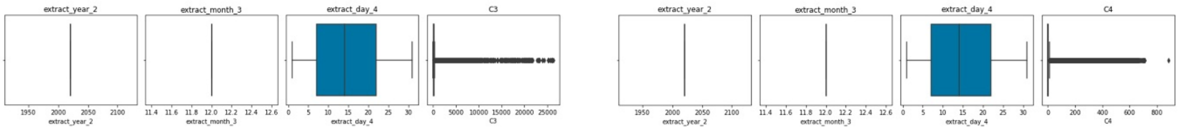


Figure 6 – Boxplots of numeric variables in non-roaming (left) and roaming (right) December dataset

In the exploratory analysis, we examined and visualize some statistical analyses of our datasets. In A 1 shows the boxplots visualizations for our numeric variables - extract_year_2, extract_month_3, extract_day_4, and C3/C4 - for all non-roaming and roaming months' datasets, in which there were no outliers considered in any dataset.

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.

Next, vodafone_grelha.xls dataset with geoinformation was imported, checked for nulls, and a boxplot generated for its numerical variables. No nulls or outliers were identified during these steps (Figure 7).

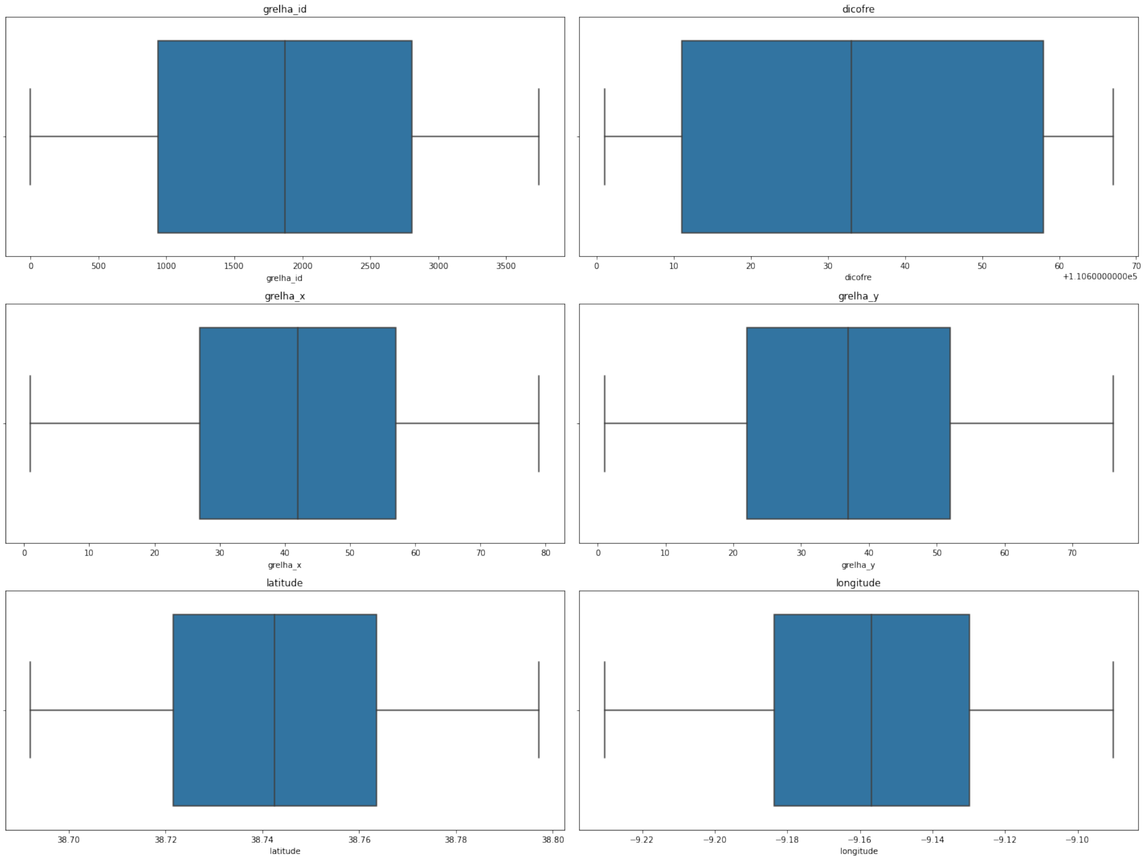


Figure 7 – Boxplots of numeric variables in September dataset

The preparation of the relevant month’s dataset to merge with the geoinformation dataset required to rename the Grid_ID for the subsequent month's dataset to grid _id as the geoinformation dataset's id and round the values of variables C3/C4 to integer. After processing the data from datasets, the columns were merged based on their identifiers (grid_id), resulting in a new dataframe.

For improved visualizations of the datasets, discretization and categorization were performed. For instance, a categorical variable was generated to identify the day of the week of the datasets’ recordings and a discrete variable was constructed to identify the time of a given record.

After examining the columns of the merged dataset, we chose to develop a data model with the columns "nome", "geometry", "datetime", "freguesia", "lat", "lon", "devices", "dia_semana", and "horas" (Table 4).

The listing of the column "nome" values was visually analyzed to remove highways from the datasets, as these locations were prone to congestion, and lead to data misinterpretation. Therefore, the following road routes were removed from the "name" column: "A5", "Eixo Norte-Sul", "CRIL", "2ª Circular", and "A2".

The multipolygon object in the geometry column needed a manual replacement of the curly braces to parentheses in order to load the geometry correctly to well-known text (WKT).

Table 4 – Table of resulted monthly datasets

Column	Description
nome	Location name
geometry	Geometric multipolygon
datetime	Object module that supplies classes for manipulating dates and times
freguesia	Parish names after new policies
lat	Latitude value
lon	Longitude value
devices	No. of distinct terminals remaining in the grid
dia_semana	No. of weekday
horas	No. of corresponding hour

We visualized the number of devices, using choropleth mapping with Folium library [26], followed by the same kind of visualization with the number of historical POIs, the number of transportation POIs (bus stop, metro station, train station) in the Lisbon parishes. The selected POIs were related to tourism and sightseeing, a combination of historical landmarks and transportation. These POIs were chosen for this analysis to understand how people moved in the city, as well as, to narrow our data model to an outstanding parish to use as the case study.

The POIs were extracted from the OSMnx library [27] by category-specific queries. We created two queries: for the historical POIs that included museums, memorials, monuments, statues, castles, castle walls, forts, city walls, churches, and bridges; for transportation POIs, first with bus stops and metro stations (POIs), followed by train station (POIs), that had to be filtered from duplicated metro station POIs.

The shapefile 'Limites_geo.shp' was loaded with the parishes of Portugal, filtered to display the parishes of Lisbon, and some parishes with accented character encoding were renamed. With the aid of the DivIcon library and HyperText Markup Language (HTML), the attributes were modified, and each parish was labelled in a more comprehensible geographical reference.

A script was developed to group the data for a given month by parish and the number of devices were added to each parish. The result of this analysis determined which Lisbon parish was selected to be our case study.

The threshold scale assigned in Figure 8 was customized based on the maximum and minimum values of the number of devices for non-roaming data for all five months. The threshold scale in Figure 8 and A 2 for non-roaming ranged from 31 million (September) to around 380 million (October), and for roaming the threshold scale ranged from 0.59 million (September) to around 33 million (November).

In December the non-roaming map showed that the core and north parishes of Lisbon had more devices. Avenidas Novas and Alvalade, for example, had more than 310 million devices, probably due to more traffic, workplaces, universities, and cultural locations. On the other hand, the roaming map shows that the inner core of Lisbon, from Avenidas Novas to Santa Maria Maior, had devices. Santa Maria Maior followed with more than 26 million devices, then Misericórdia, Santa António, Avenidas Novas, and Olivais, with more than 13 million. Estrela and Arroios had around 7 million, and the remaining parishes fewer than 7 million devices.

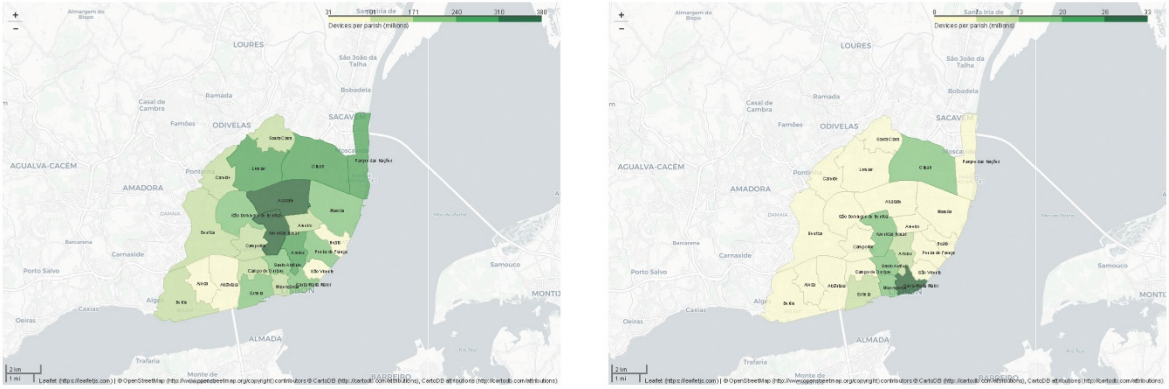


Figure 8 – Devices (millions) per parish in December non-roaming (left) and roaming (right) dataset

In A 2, we categorized the data by month in non-roaming and roaming, the first column with non-roaming data and the second column with roaming data. This figure showed a gradual increase of devices for each month. Note that in November, there was no available data for non-roaming or roaming in Santa Clara parish.

From a broad perspective, recalling what was previously discussed in Figure 8, we observed an increase in the number of devices over the months in the inner core of Lisbon, from Alvalade to Santa Maria Maior.

Due to the Humberto Delgado's airport flow, the non-roaming maps had exceptional high number of devices in Olivais. It stood out in every month, with values at or above 13 million devices, with the exception of November, which has the lowest values, with fewer than 7 million devices.

Avenidas Novas, non-roaming data presented more devices with more than 310 million monthly, whereas Beato and Ajuda presented the lowest values of the scale with less than 101 million monthly. On other hand, for roaming data, Santa Maria Maior was the parish with the highest number of devices in all five months (13 million in September, and 26 million in October, November, and December, and more than 20 million in January).

On the roaming and non-roaming maps, in almost all parishes, there was a device increase from September to October, as well as in the following months.

Next, we performed the analysis and visualization of the number of POIs per parish: first, the number of historical POIs, followed by transportation POIs - bus stop, metro station, train station.

In Figure 9 showed the absence of historical POIs in the parishes of Santa Clara, Beato, and Penha de Franca. On the other hand, Santa Maria Maior had more than 53 historical POIs, followed by Misericórdia (42 historical POIs) and Santo António (22 historical POIs). The other parishes had na average of twelve historical POIs. Based on the historical POIs map visualization (choropleth), we understood that Santa Maria Maior was the Lisbon parish with the most significant number of historical POIs, mainly due to its heritage buildings and characteristics.

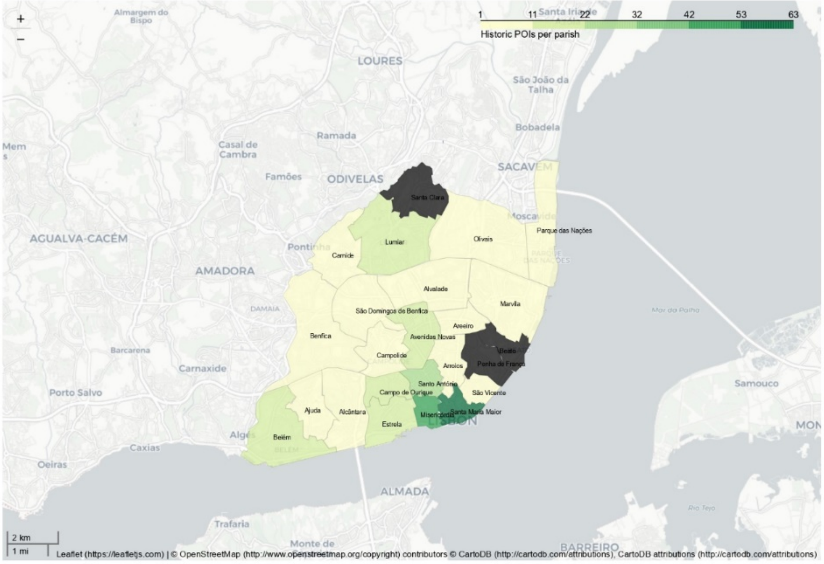


Figure 9 – Historic POIs per parish

Lumiar and Marvila parishes stood out (Figure 10) with more than 142 bus stops, in each, followed by Olivais and Benfica, both with more than 115 bus stations. The highest number of bus stops were located in the northern parished, indicating a strong bus infrastructure in parishes with residential only characteristics.

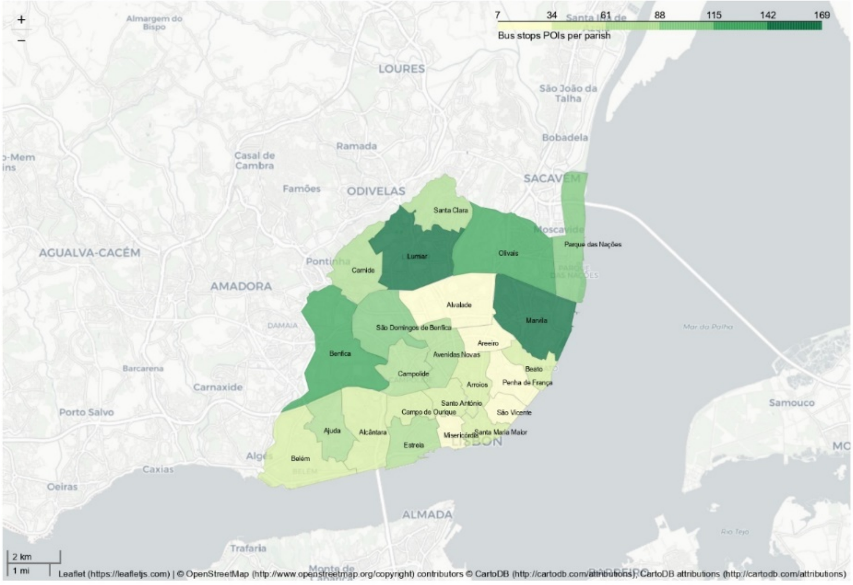


Figure 10 – Bus stop POIs per parish

In Figure 11 we observed that metro stations did not exist in the parishes of Belém, Ajuda, Alcântara, Campo de Ourique, Estrela, São Vicente, Penha de Franca, and Beato. Santa Maria Maior and Avenidas Novas had the greatest number of metro stations, six, followed by Olivais and Lumiar with more than four stations and Santo António e São Domingos de Benfica with more than three stations.

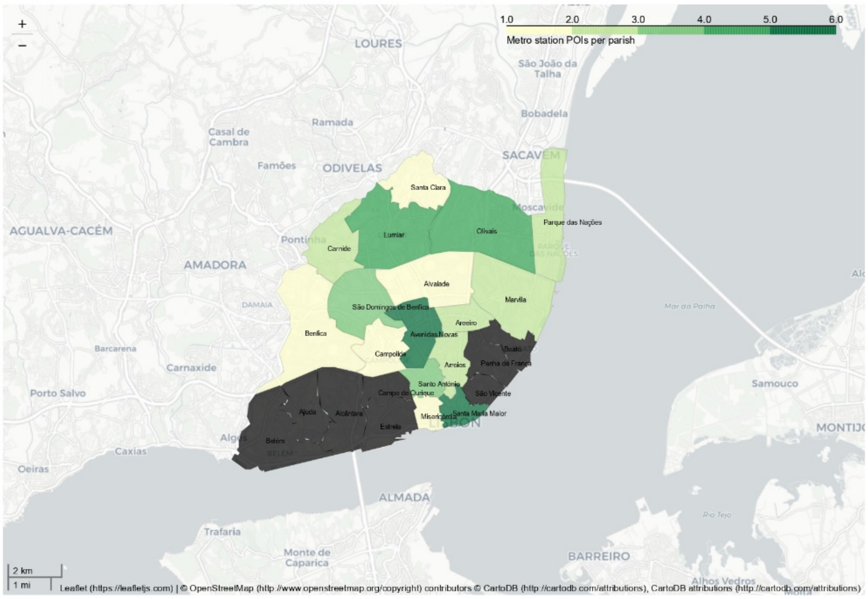


Figure 11 – Metro station POIs per parish

In Figure 12, the parishes greatest number of train stations were in Estrela, Campolide, Alvalade, and Marvila, each had three train stations. Next, Parque das Nações, Benfica, Belém, São

Domingos de Benfica, Misericórdia and Santa Maria Maior, with two train stations. The remaining fourteen parishes lack train stations.

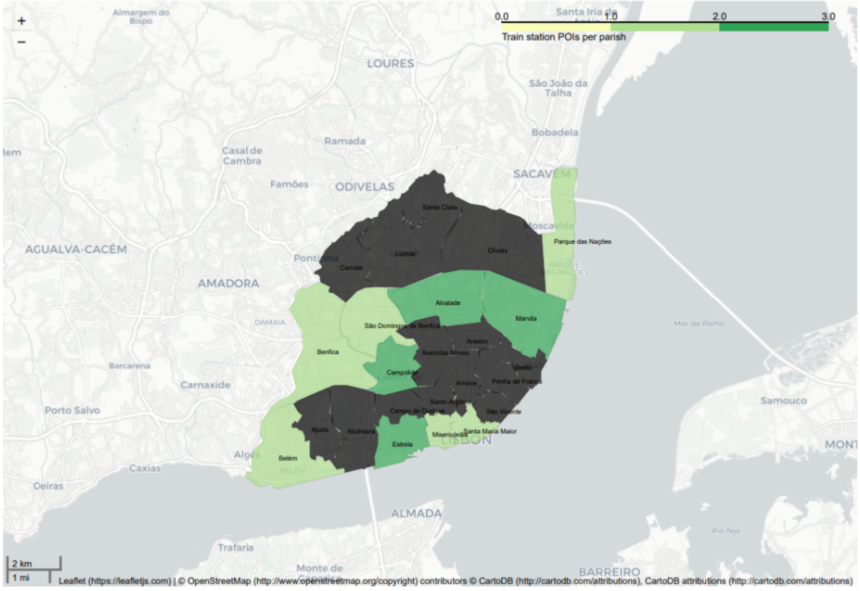


Figure 12 – Train station POIs per parish

According to of the analysis and visualization in Figure 9, Figure 11 and Figure 12, the parishes of Santa Maria Maior and Marvila stood out with the greatest number of transportation POIs, metro and train stations, and bus stops combined. Furthermore, Santa Maria Maior was also the parish with largest number of historical POIs, as such, for further analysis in our thesis, Santa Maria Maior was selected as case study for our data modeling. This decision was also supported by the fact that Santa Maria Maior (Figure 8) counted on with a large number of devices (between 31 million and 101 million).

3.4. Data modeling

In this section, we present the data modeling results regarding Santa Maria Maior by analyzing the devices data with POIs data, presenting insights on people's mobility patterns in Lisbon. We analyzed all the months (see Appendix) but in this section we only present the month of December.

First, we analyzed POIs data, followed by devices data, and a combined analysis of both. Finally, we applied DBSCAN, resulting method of our SLR, and performed a forecasting algorithm to specific clusters of our DBSCAN results.

3.4.1. Smartphone data analysis and visualization

We analyzed the smartphones' data to understand users' behavior over time in Santa Maria Maior, looking at weekly (Figure 13), hourly (Figure 14), and comparing weekdays and weekends ( ). As described in section 3.3 , we added two columns 'dia_semana' and 'horas' (Table 4), by categorizing the days of the week from the year, month, and day columns, and discretizing the hours by the 24 hours of the day from our Datetime column.

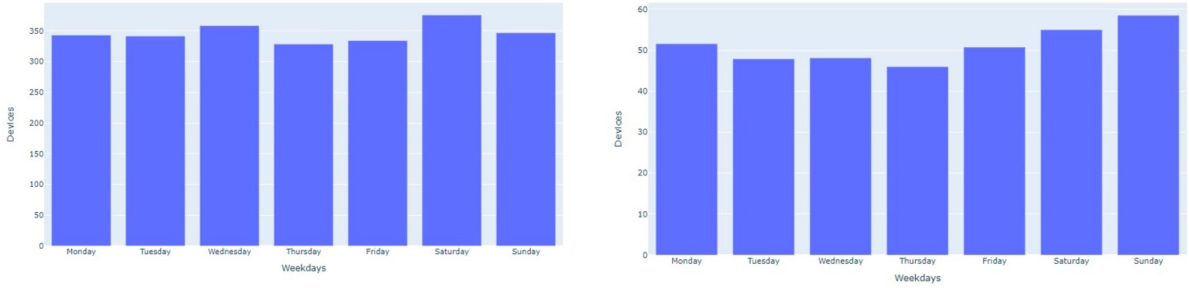


Figure 13 – Average devices (thousands) over December's non-roaming (left) and roaming (right) weekdays in Santa Maria Maior

In Figure 13, we observed in the non-roaming bar plot that the number of devices tended to be very similar from Monday to Tuesday (350 thousand devices), rising on Wednesday (over 350 thousand devices), with a slight drop on Thursday (325 thousand devices), and rising until Saturday (375 devices), and then dropping on Sunday. We observed a weekdays average of 350 thousand devices.

In the roaming case, the weekly behavior tended to decrease on Monday (50 thousand devices), followed by Tuesday, Wednesday and Thursday (under 50 thousand devices), and increasing from Friday (50 thousand devices) onwards, peaking on Sunday with nearly 60 thousand devices.

Overall, all months showed in non-roaming, a pattern of growth from Monday to Friday, extending to Saturday in the months of October and January, apart from December, which maintains constant values throughout the week. As for all months in the roaming case, showed a growth pattern throughout the week, except for October (decreased from Sunday to Wednesday and rose until Saturday). This showed that the Portuguese population had a regular mobile phone use throughout the week, while foreigners tended to be more active on Fridays and weekends. In A 3 all visualizations are found for the devices average by weekdays for non-roaming and roaming for the five months.

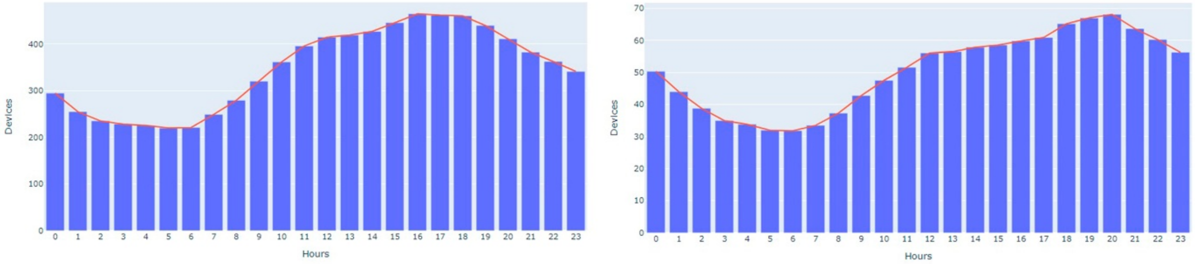


Figure 14 – Average devices (thousands) over December's non-roaming (left) and roaming (right) daily hours

Figure 14 showed the average number of devices, non-roaming and roaming, by time of the day. In non-roaming we observed the lowest activity at 6 a.m. (slightly more than 200 thousand devices), starting to grow and reaching the maximum peak at 4 p.m. (around 450 thousand devices), taking 10 hours. This daily pattern is correlated with the commute times, people go to work or study between 8 a.m. and 9 a.m. and go back home between 4 p.m. or 7 p.m. This explains the influx of devices during day and, consequently, the decline in devices after this time.

Overall, the number of devices in day peak non-roaming, range from 350 thousand in January and 375 thousand in October, and 450 thousand in November and December. For roaming data, we observed the lowest activity at 6 a.m. (little more than 30 thousand devices), growing till the peak at 8 p.m. (70 thousand devices), requiring 14 hours. A 4 contains all graphs corresponding to the average number of non-roaming and roaming devices for all the months.

In Figure 15 we compared 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).

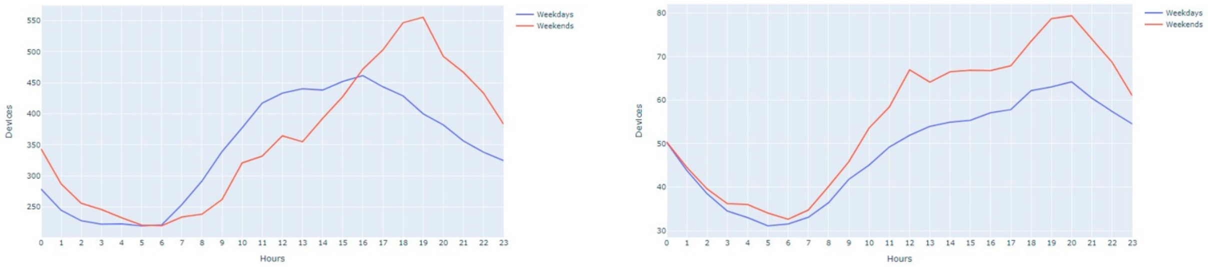


Figure 15 – Average devices (thousands) for weekdays (blue line) and weekends (red line) over December's non-roaming (left) and roaming (right) daily hours

The non-roaming graph showed a similar pattern to Figure 14. In the weekdays and weekends the number of devices started increasing at 6 a.m. till the peak at 4 p.m. and 5 p.m. Furthermore, Santa Maria Maior had more devices in the weekdays during the day and in the weekends during the night.

The devices in the roaming graph showed similar patterns in the weekdays and weekends, despite the highest number of devices in the weekend explained by the big inflow of tourists in this parish. In the weekdays, the number of devices began to increase at 5 a.m. and continued to rise until 8 p.m., at which point they began to decrease until 5 a.m. In the weekends started increasing at 6 a.m., peaking at 12 p.m., declining until 1 p.m., and rising again until 8 p.m. declining until 5 a.m.

In all the analyzed months, the non-roaming weekdays and weekends showed the same pattern. In September and October, there are more devices in the weekends between 7 p.m. and 6 a.m., in November and December between 4 p.m. and ends at 7 a.m., in January between 5 p.m. to 7 a.m. Only in the month of December, the weekends peaked higher than the weekdays. We observed a rise from just over 400 thousand devices in October to 450 thousand devices in November, and 550 thousand devices in December, followed by a decline to under 400 thousand devices in January. For the roaming months scenario, the pattern was similar, with a few exceptions in September, where the average number of devices for weekend roaming users is consistently higher. Looking at the highest values for each month, September was a little over 80 thousand devices, October increased to nearly 90 thousand devices, November continued its ascent to 110 thousand devices, and in December began to decline to around 80 thousand devices, following to January with 55 thousand devices. A 5 contains all graphs corresponding to the monthly average of non-roaming and roaming devices weekdays and weekends.

3.4.2. Smartphone data and POIs analysis and visualization

We associated the number of devices in the square to our data model with different POIs categories, i.e., for each month a script: if a given geometry point of a POI is within a given polygon of a quadrant, that POI called the number of devices for that quadrant. Thus, it was understood that ‘x’ number of devices had remained more than five minutes at specific POI.

The same approach was developed to define the threshold scale used in section 3.3, used in the following graphs, i.e., the maximum and minimum values for non-roaming and roaming were determined for each month. The threshold scale obtained based on the quadrants with non-roaming data from the parish of Santa Maria Maior had a minimum overall value of 0 million devices, since the minimum value detected was 0, and a maximum overall value of 9.8 million devices. For the roaming data the threshold scale had a minimum of 0 and a maximum of 1.8 million devices.

As there were few metro and train POIs, we chose to combine them in the analysis.

In Figure 16, Figure 17 and Figure 18 for both non-roaming and roaming data, it is showed a high concentration in the west and core of Santa Maria Maior, with the north of the non-roaming map highlighted. Both maps show a high number of devices in the Chiado.

In Figure 16 and Figure 17, there were orphan POIs, not considered, as they were located beyond the quadrants of Santa Maria Maior. These POIs were excluded from future analysis as their point geometry lacks a quadrant polygon.

Figure 16 and Figure 17 showed a dense concentration of bus stops (POIs) and historic POIs all over Santa Maria Maior. There were four orphan POIs for bus stops and six orphan for historic POIs that were not taken in account. Finally, Figure 18 displayed the railway stations (POIs), which were spread along Santa Justa, Socorro, Chiado, and Madalena.

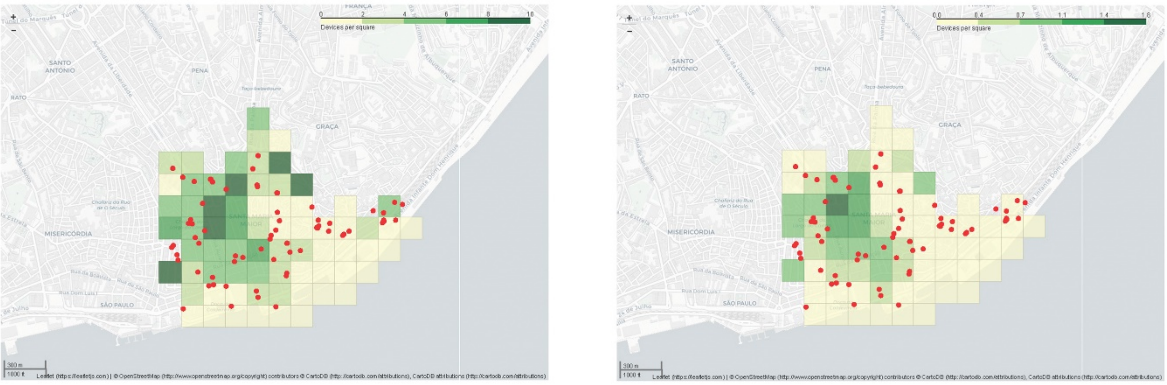


Figure 16 – Historic POIs influence area (red dots) in Santa Maria Maior's smartphone data quadrants in December non-roaming (left) and roaming (right)

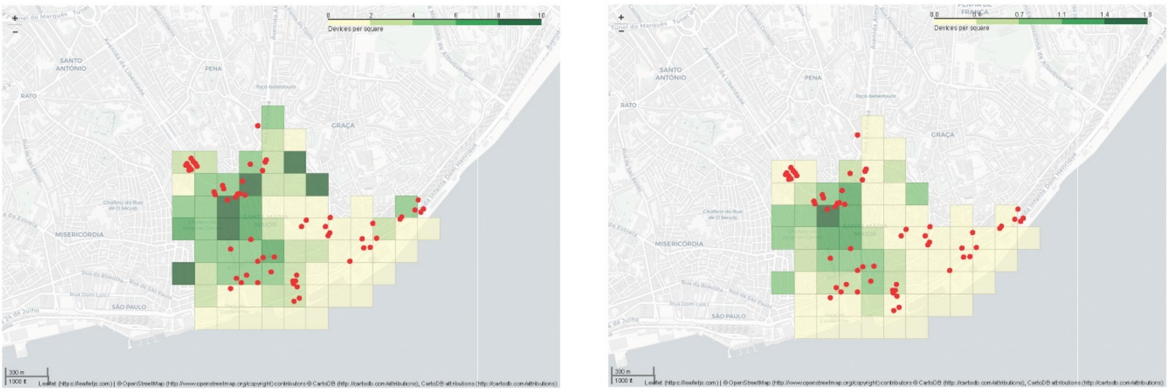


Figure 17 – Bus stop POIs influence area in Santa Maria Maior's quadrants in December non-roaming (left) and roaming (right)

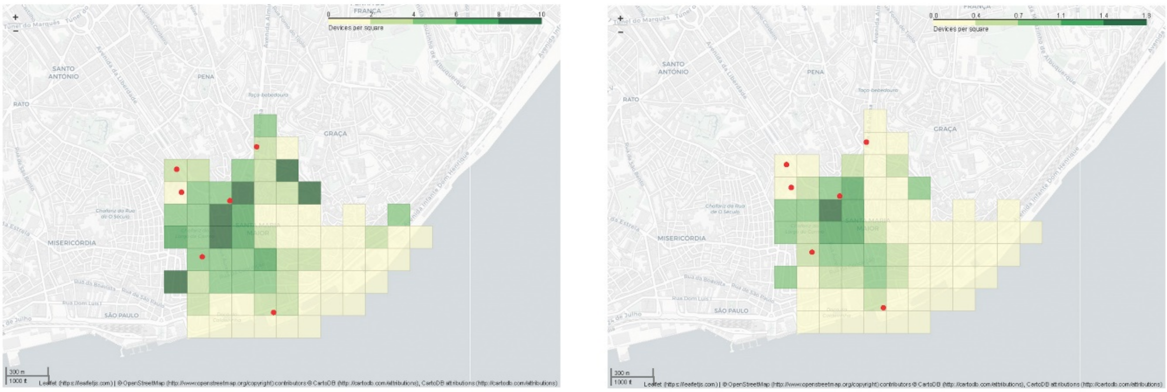


Figure 18 – Metro and train POIs influence area in Santa Maria Maior's quadrants in December non-roaming (left) and roaming (right)

In A 6, A 7 or A 8, the easternmost of Santa Maria Maior lost intensity from October to December, regardless of roaming or non-roaming scenario. The September’s low device density was largely related to the dataset's data limitation, as well as, January for reasons the season lower temperatures and consequent fewer tourists (roaming that does not exceed 1.1 million devices).

In A 6 and A 7, we observed that October, which is the month that fills quadrants with values greater than 2 million for non-roaming data, and densities greater than 0.4 million for roaming data where the majority of historical and bus stop POIs are inserted. We observed in A 8 with the exception of the month of September for non-roaming and roaming, which lacks data, the points of interest of the railway transports fit quite well in the occupation of the quadrants with high device values.

Figure 19 depicted the histogram of the total number of POIs. The 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 Figure 19, we observed that the category with the biggest number of POIs in Santa Maria Maior were the historical POIs, while the category with the lowest number were the train stations. Following the historical POIs with 62 points are bus stops with 24 points, metro stations with 5 points, and train stations with 1 point. As mentioned before, this parish had a high concentration of historical POIs

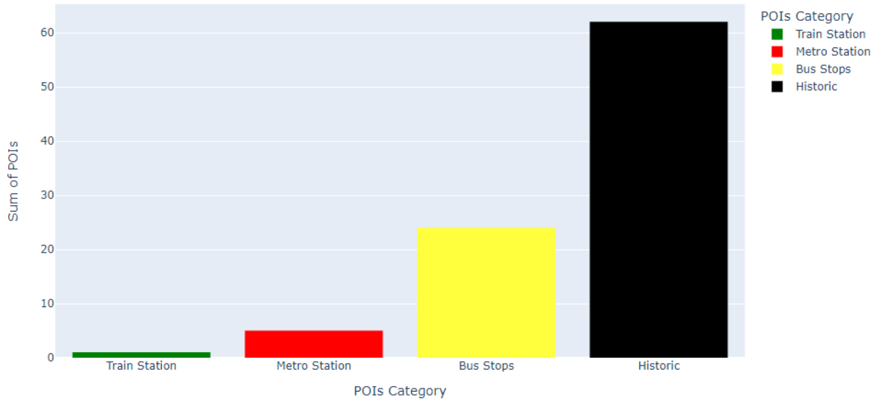


Figure 19 – POIs category histogram

In Figure 20, we mapped the POIs by categories - historical, bus stops, metro stations, and train stations - into a scatter plot with two bar charts for: the x-axis representing longitude and the y-axis latitude. In the POIs regarding public transportation, we verified strategic positioning throughout the parish of Santa Maria Maior. We may even claim that there was a transportation combo due to the multiple connections, particularly in Rossio in the Santa Justa, where there were nine bus stops, one metro station, and one train station. The presence of bus stops and historic POIs across the parish of Santa Maria Maior was notable. We observed the maximum concentration of POIs between latitude 38.710 and 38.714, with slightly over 40. The other POIs were scattered along Santo Estêvão, São Miguel, Alfama, Santiago, Sacramento, Chiado, and Castelo. The highest concentration in longitude, between -9.1425 and -9.1400, with only 20 POIs, were mostly bus stops and historical places, followed by metro stations and the train station scattered in Santa Justa, Sacramento, Chiado, Mártires, and São Nicolau, as well as other bus stops and historical POIs in Castelo, Santiago, and Sé.

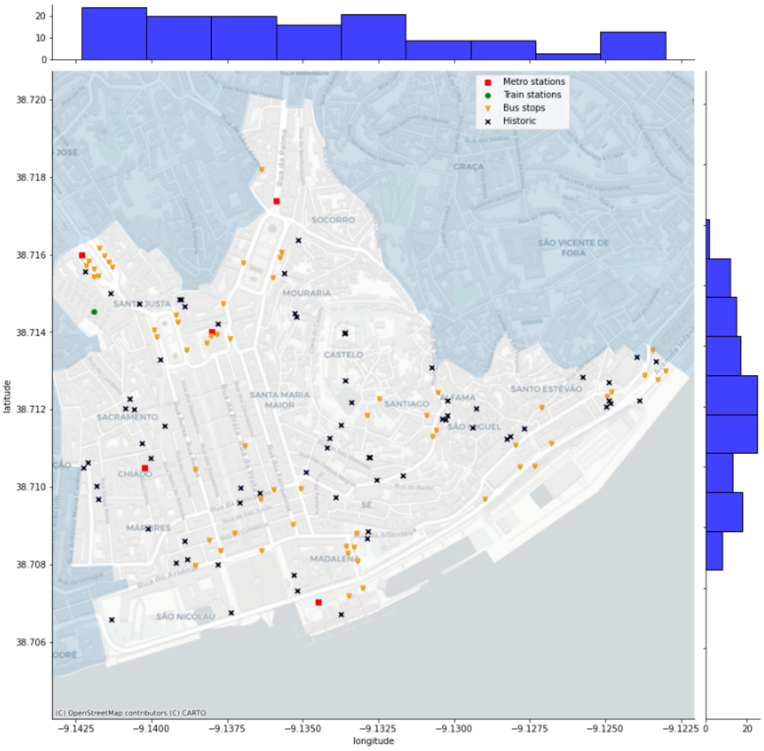


Figure 20 – POIs location distribution in Santa Maria Maior

A more detailed POIs category analysis was performed, and looked at month by month, on how many devices there were in every POI.

To analyze the historical POIs we made a script to reorganize the original POIs groups — castle, memorial, monument, etc. — and created four groups: castle, memorial, monument + museum and others with the remaining POIs.

Figure 21 provided an overview of the four types of historical POIs that were studied for non-roaming (left) and roaming (right) devices in December. The memorials had the highest number with 44, followed by castles with 10, others with 4, and monuments with 2. A later-created type called "others" comprised churches, city walls, and archaeological sites.

In the non-roaming and roaming the historical POIs, the memorial had the greatest number of POIs, as well as the greatest number of devices, with more than 2.5 million (mean value) devices for non-roaming, and 3 million devices for roaming, more than half of its POIs. The memorial of the Incêndio do Chiado had nearly 5.5 million devices for non-roaming, and 7.5 million devices for roaming, and the memorial of Maria José Nogueira Pinto and Comemoração do Navio-Escola Sagres were the sites with the fewest number, with less than 1.5 million devices for non-roaming and roaming.

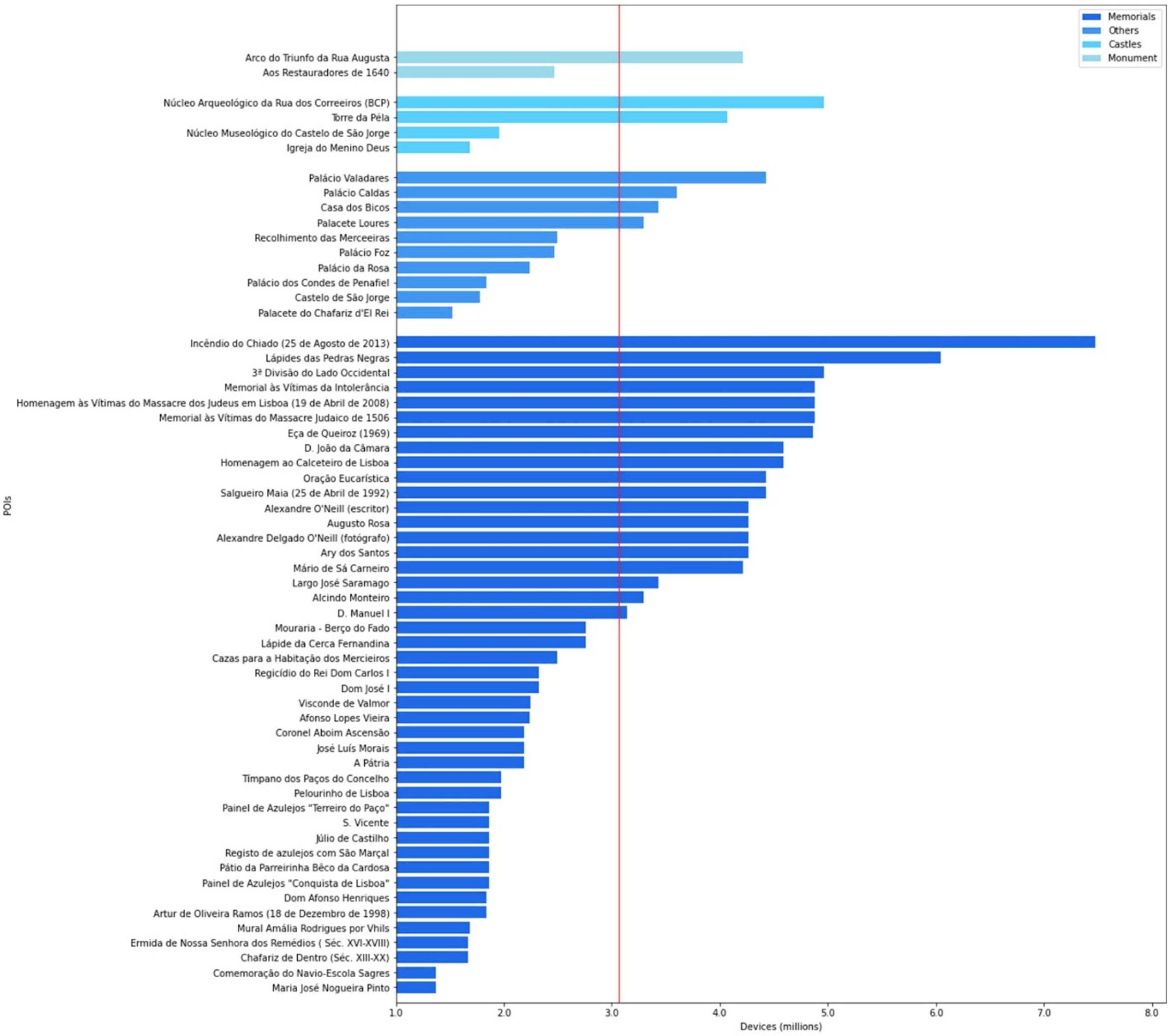


Figure 21 – Santa Maria Maior historic POIs with non-roaming devices' average (red line) in December roaming

In A 11, we showed the five months analysis for non-roaming and roaming for the historic POIs, where it stood out the predominant number of devices in the memorials, followed by the others, castles and monuments, highlighting the month of December for roaming. In all months, the memorials POIs consistently displayed nearly half of the total. In the non-roaming, the average number of devices decreased significantly from almost 4 million in September to about 2.5 million in December, and then rose to about 3 million in January. On the other hand, the data for the roaming months indicate a significant increase in their average from September with nearly 2.5 million to November with approximately 3.75 million, followed by a decline to just over 3 million in December and just under 2.5 million in January.

Figure 22 showed non-roaming and roaming devices included in the bus stops POIs. In the non-roaming, Praça da Figueira had nearly 35 million devices, Martim Moniz over 25 million, and Praça do Comércio nearly 25 million devices. The greatest number of devices in bus stops POIs was found in December. Where areas, Jardim do Tabaco, Rua dos Remédios and Chafariz de Dentro had the fewest devices, totaling over 1 million.

In the roaming, the top three areas of devices in bus stop POIs were the same, with an order shift between second and third: Praça da Figueira with more than 5 million devices, Praça do Comércio with over 3.5 million, and Martim Moniz with over 3 million. Also, the bus stop POIs with the lowest number of devices were Jardim do Tabaco, Chafariz de Dentro, and Rua dos Remédios, which totaled less than half a million.

When comparing non-roaming and roaming results, the difference was apparent, as the non-roaming had an average of around 8 million and the roaming case had an average of approximately 1 million.

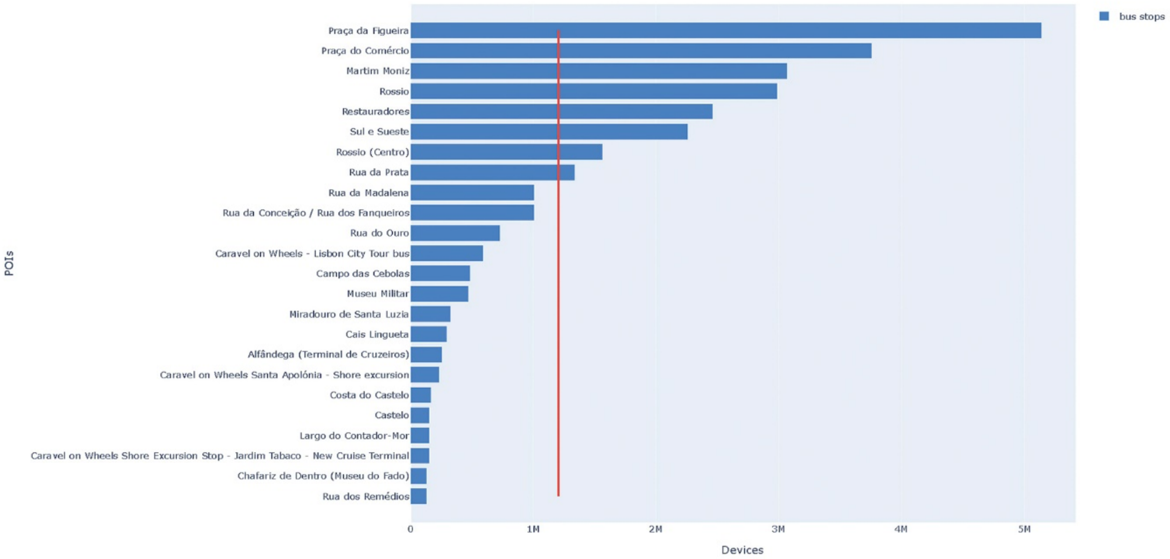


Figure 22 – Santa Maria Maior bus stops POIs with devices' average (red line) in December roaming

In A 9, whether non-roaming or roaming, all the months showed the same pattern. Approximately eight bus stop had devices above the mean and the remaining ones had values below the mean. Praça do Comércio, Praça da Figueira, Restauradores and Martim Moniz were the bus stop POIs with the highest number of devices. Rua dos Remédios, Jardim do Tabaco, Chafariz de Dentro, Castelo, and Largo do Contador-Mor were the areas with the smallest number of devices in bus stops POIs. Overall, for non-roaming, the number of devices in bus stop POIs increased from October to December before decreasing in January, between 5 to 10 million devices. In the non-roaming the number of devices increased above the average from October to November, declining in subsequent months. The decrease pattern was related to the months of winter and to the pandemic restrictions.

The metro and train station POIs were grouped together for the same rationale as previously presented. In Figure 23 contained the POIs corresponding, displayed separately and in ascending order. The railway station Rossio had the greatest number of devices in non-roaming and roaming. The metro station with the greatest number of devices in non-roaming and roaming was Restauradores in December with more than 7 million devices and 900 thousand devices, while the fewest devices were found in Terreiro do Paço metro station in non-roaming (3 million devices) and Martim Moniz metro station in roaming (close to 400 thousand devices).

In a comparative study between non-roaming and roaming, the difference in average values was easily visible, as in non-roaming had an average of around 4.5 million and in roaming had an average of approximately 500 thousand devices.

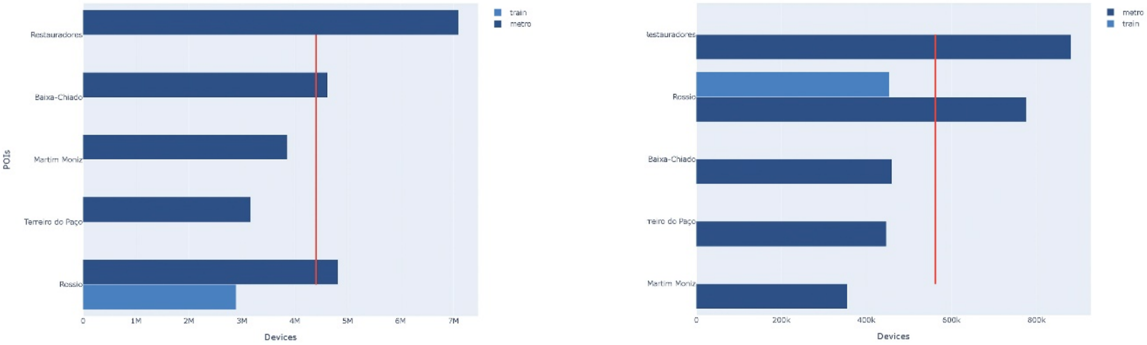


Figure 23 – Santa Maria Maior train and metro stations POIs with devices' average (red line) in December non-roaming (left) and roaming (right)

In A 10, we performed a five months analysis for non-roaming and roaming, and we observed in the non-roaming that Restauradores stood out with the highest number of devices in every month, with Baixa Chiado also with a number of devices above the average, except in December, when Rossio exceeded the average value. In roaming, Restauradores also outstood, from September to October. In the following months, Rossio surpassed the norm, while Martim Moniz fell below.

In the non-roaming case, the average begins in September with approximately 1.25 million devices and peaks in December with nearly 4.5 million, whereas in the roaming case, the average begins in September with approximately 255 thousand devices, peaks in November with nearly 700 thousand devices, and declines to approximately 400 thousand in January.

3.4.3. Smartphone data and POIs clustering

DBSCAN collects points that are close to one another based on a Euclidian distance measurement and the smallest number of points based on a collection of points. It also identifies outliers in low-density areas.

As in Figure 24 this DBSCAN has three types of data points. If a point has more than MinPts points within eps (ε), it is designated a core point. A point within eps (ε) with fewer than MinPts that is close to a core point, it is a border point. A point that is not a core or border point is referred to as a noise point or outlier.

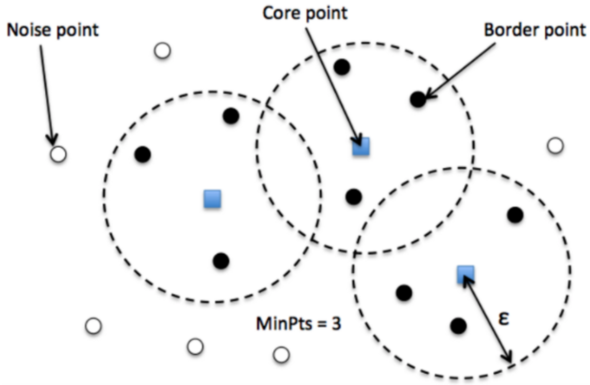


Figure 24 – Diagram of DBSCAN point types

Our goal was to identify correlations and structures in data that would be difficult to find manually but could be essential and useful in recognizing patterns and anticipating trends in the POIs of our case study in Santa Maria Maior.

First, we conducted separate DBSCAN analyses for the bus stops and historical POIs. Due to the low density illustrated in Figure 18 in section 3.4.2, we did not consider metro and train stations POIs. In order to increase the density of the DBSCAN analysis, we performed the bus stops and historical POIs collectively.

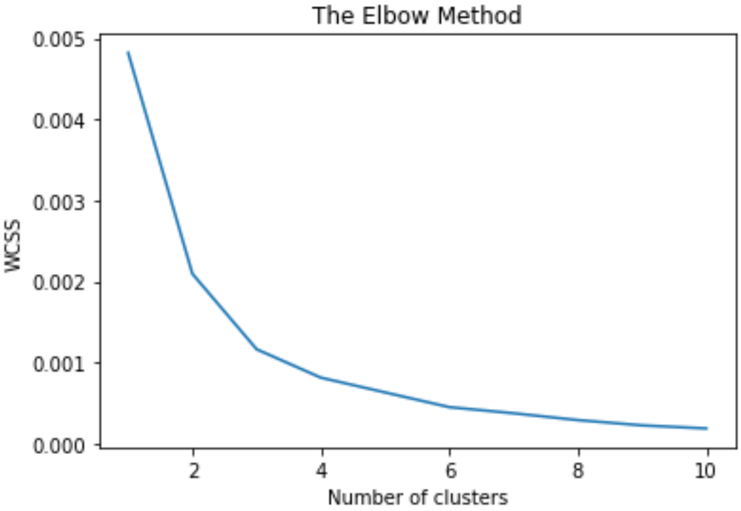


Figure 25 – The Elbow method for historic, bus stops, metro, and train stations POIs

To better analyze and contextualize the generated clusters, we constructed a scatter plot with the parish boundaries of Santa Maria Maior to a visualize the correlation with the devices by quadrant. In section 3.4.2, we applied a technique for determining the number of devices per quadrant that we reused for a more accurate interpretation. These maps were produced for each of the five months for both non-roaming and roaming data.

To select the optimal value for the DBSCAN epsilon, we applied the elbow method (Figure 25), generating a plot with k-distance that included the number of iterations from 1 to 11 for the K-Means algorithm. As a result, average distances on the y-axis and the cluster points on the x-axis.

In the following figures generated for the month of December, red dots represented the outlier POIs (these were discarded during clustering).

In Figure 26, based on the elbow method (A 2, first graph), epsilon was set in 0.004 for the historical POIs, cluster sample of four that generated six clusters and accounted for forty nine outliers in a total of sixty nine POIs. In Figure 27, applied the same method for bus stop POIs (A 2, second graph), with epsilon of 0.0004, cluster sample of three, that generated eight clusters and accounted for forty three outliers in a total of sixty POIs.

In Figure 28, showed the same analysis for both historical and bus stop POIs non-roaming (A 2, third graph), with epsilon set to 0.001, cluster sample of three that generated eight clusters and accounted for sixteen outliers in a total of hundred and twenty nine POIs. In Figure 29 the analysis was performed with the same POIs as before and same parameters (epsilon 0.001 and cluster sample of three) for roaming (A 2), resulting in eight clusters as well, and fifteen outliers in a total of hundred and thirty five POIs.

In Figure 26 and Figure 27 are illustrated the worst-case DBSCAN clustering scenarios. Figure 27 revealed several outliers with minor clusters in Sacramento, Santa Justa in Largo São Domingos, Mártires in Praça do Município, São Miguel and Santo Estêvão. In Figure 27, there were also small clusters and many outliers, namely in Santa Justa, Praça dos Restauradores and Praça da Figueira, and in Madalena, in Rua dos Amareiros. In Figure 28 and Figure 29 showed clusters with the fewest outliers: Santa Justa, the black cluster, along Praça dos Restauradores, Praça Dom Pedro IV, and Praça da Figueira; Mouraria, pink cluster, in Praça Martim Moniz; Chiado and Sacramento the white cluster; the longest cluster, in purple, from Mártires to Castelo and Santiago; the yellow cluster in Madalena; the brown cluster around Sé; the orange cluster included Alfama and São Miguel; the blue cluster in São Miguel and São Estêvão; and finally, close to São Vicente parish, Santo Estêvão area - around the Lisbon Military Museum - the green cluster.

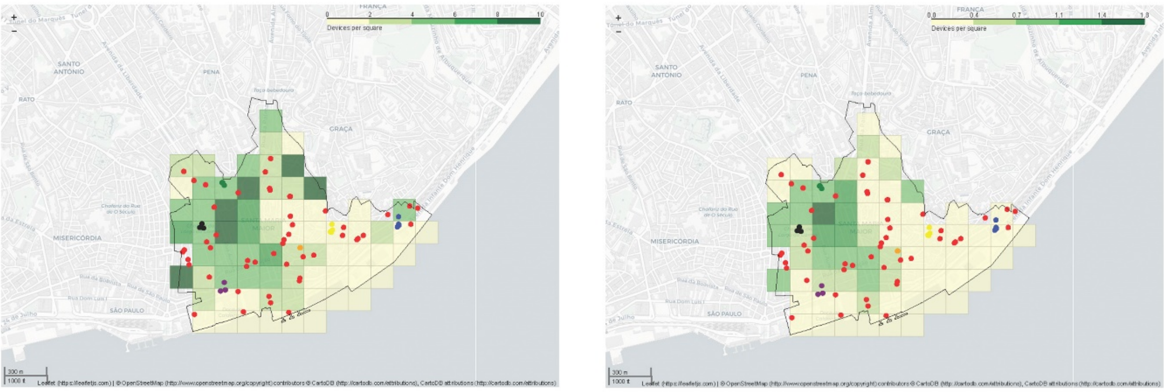


Figure 26 – DBSCAN historic POIs influence area (colored dots) in Santa Maria Maior's smartphone December's non-roaming (left) and roaming (right) data quadrants (millions)

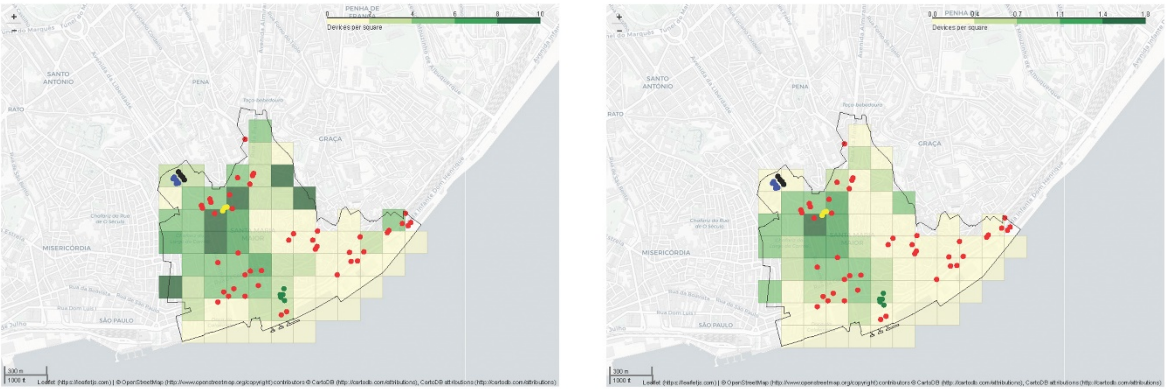


Figure 27 – DBSCAN bus stops POIs influence area (colored dots) in Santa Maria Maior's smartphone December's non-roaming (left) and roaming (right) data quadrants (millions)

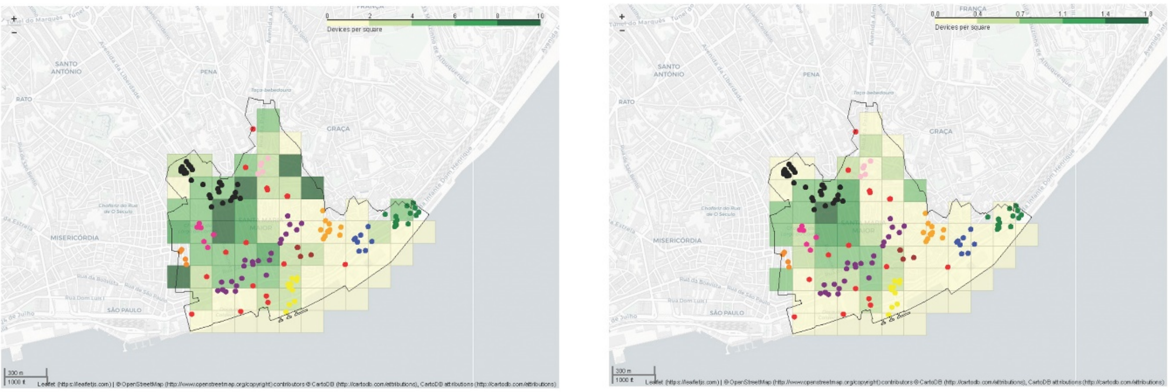


Figure 28 – DBSCAN historic and bus stops POIs influence area (colored dots) in Santa Maria Maior's smartphone December's non-roaming (left) and roaming (right) data quadrants (millions)

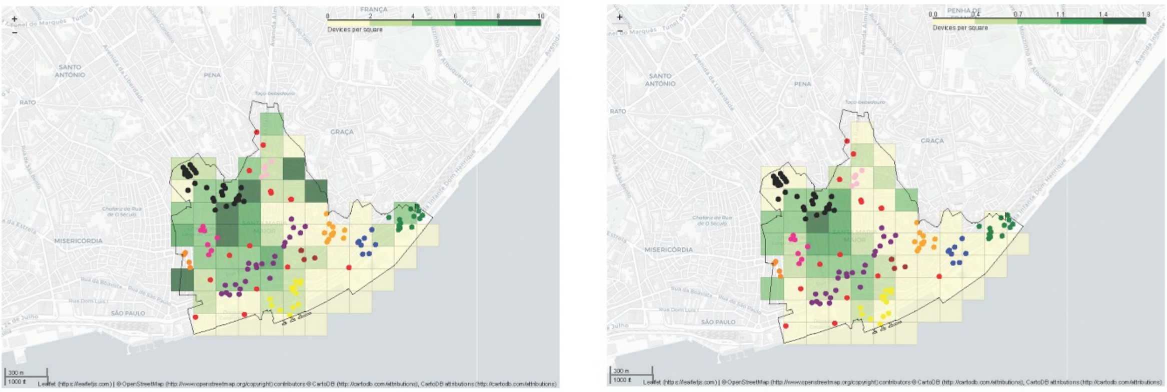


Figure 29 – DBSCAN historic, bus stops, metro, and train stations POIs influence area (colored dots) in Santa Maria Maior's smartphone December's non-roaming (left) and roaming (right) data quadrants (millions)

.

In A 16, all five months, non-roaming and roaming were considered. In non-roaming, most clusters were located in the west and in the core of Santa Maria Maior, including the black, pink, dark pink, purple, and yellow clusters, with the exception of the green cluster, in the east, with 6 to 8 million devices. The clusters found in the roaming scenario were identical with those in the non-roaming data, with a smaller number of devices that decreased in eastern area.

As a result of the larger number of POIs and devices the DBSCAN generated more compact clusters with fewer outliers. Of the resulting eight clusters, we highlight two clusters, the purple and the blue. The purple cluster located along the well-known electric tram 28E | Martim Moniz – Prazeres route, crossing Baixa area, and the blue cluster located around the Lisbon Cruise Terminal. Both cluster locations were in the downtown area with tourist attractions although the purple cluster for its transportation POIs also had a high density of non-roaming and roaming.

3.4.4. Forecasting with the Prophet algorithm

We opted to conduct a forecast study in the two outstanding clusters - purple and blue – identified in the DBSCAN analysis. We performed a forecast model to understand the non-roaming and roaming mobility patterns.

The practice of evaluating time series data with statistics and modelling to create forecasts and provide information for strategic decision-making is referred to as time series forecasting. It is not always possible to make an accurate prediction, and the probability of forecasts might vary widely. Forecasting lets you know which possible outcomes are more likely or less likely to happen than others. Most of the time, the more information we have, the better our forecasts can be. The objective of the time series analysis is to acquire an understanding of the underlying causes of the data through the development of models. Forecast analysis can help users figure out what to do with what you know and what you can predict in the future.

In this thesis we were interested in the application of a forecast algorithm in our five months data to predict the following month, February 2022. We chose Prophet, an open-source software developed by the Core Data Science team of Facebook. It is an approach for forecasting time series data based on an additive model integrating yearly, monthly, and daily seasonality, as well as holiday impacts. Prophet is resilient to missing data and trend shifts, and it typically handles well outliers, but if we want to get the most out of this method, we need to use time series with significant seasonal effects and data from numerous seasons.

To begin with this analysis, we took into account the fact that the data to be analyzed corresponded to a time of pandemic restrictions, which altered the population's mobility behavior and, therefore, the mobility patterns. These patterns were not the regular ones and different with a decrease of users in the last of the five months of analysis. However, we believe to be interesting to create a prediction using this data.

In section 3.4.3, we considered two clusters with attractive patterns, namely the purple cluster and the blue cluster, on which we applied the forecast model, i.e. the datetime and device number data of the quadrants where these clusters' POIs were placed were considered for our new four working datasets: blue cluster non-roaming dataset, purple cluster non-roaming dataset, blue cluster roaming dataset and purple cluster roaming dataset.

We renamed the columns of all new datasets from "devices" to "y" and "datetime" to "ds" to comply with the Prophet library requirement. We generated boxplots depicting the number of devices per row for each dataset for roaming and non-roaming months, to identify outliers that could affect the algorithms' performance.

In A 17, for the non-roaming data corresponding to the blue cluster (left column of the grid) in the October, we eliminated outlier with values of more than 2000, and for the November, we eliminated outlier with values greater than 1000. And, in A 18, for the data corresponding to the purple cluster (left column of the grid), we eliminated values above 2000 from the September, above 3000 from the October, and above 3700 from the November. In A 17, in the case of roaming data for the blue cluster (right column of the grid) in the October, values larger than 200 were eliminated. In A 18, in the case of the purple cluster (right column of the grid), values greater than 700 were removed in the October, values greater than 570 were removed in the November, values greater than 700 were removed in the December, and values greater than 270 were removed in the January.

All the months of the blue cluster and the purple cluster were then concatenated separately for non-roaming and roaming. The criteria for the removal of these values was interleaved with different states with varying values, as in Figure 30, with or without the removal of values deemed to be outliers.

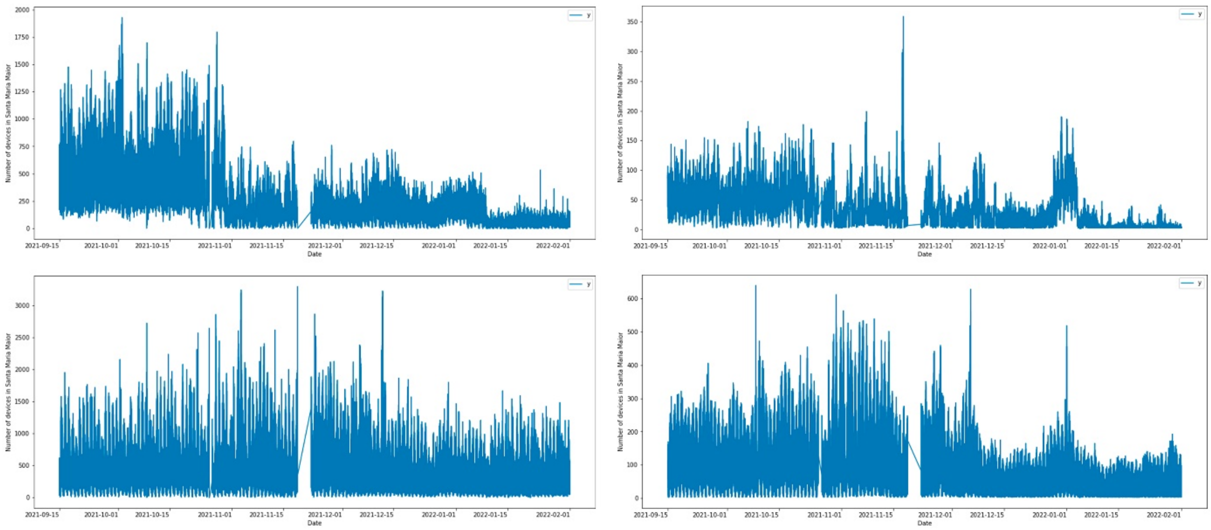


Figure 30 – Time series of blue cluster (first row) and purple cluster (second row) non-roaming (left) and roaming (right) during all months' dataset

Figure 30 with four temporal representations of the number of devices in Santa Maria Maior for our new four datasets. In the blue cluster for non-roaming, we found cyclical movement patterns ranging from zero to nearly two thousand devices over the five months. The variation and number of devices were relatively high from September 15th to the end of October, with a decline in the beginning of the following month and a continuation of the same trend until the first quarter of January, when decreased again. From September 15th to slightly more than half of November, the blue cluster with roaming exhibited a consistent trend, with values ranging from 0 to 200 devices. Following this point, there was a sharp increase in the daily dimensions between 300 and 350 values, before returning to the initial pattern with a change in its cyclical values in the end of 2021 and in the beginning of the following year. This easily reflected the strong allocation of people to this downtown area that was very festive in this time of year. In the case of the purple cluster for non-roaming the pattern over time was quite uniform, but from the September 15th until the end of November, its variance revealed steadily increasing peaks over time. Aside from the peaks, the values remained between 0 and just over or under 2000 devices. Values declined to September levels in the beginning of the second half of December.

Overall, the purple cluster for non-roaming and roaming had greater values than the blue cluster for non-roaming and roaming. In both instances, the purple cluster had a more consistent pattern than the blue cluster.

Due to the lack of data in October and November datasets, an autofill was conducted to fill them. Furthermore, the dataset was prepared for the Prophet library integration to start with the forecast analysis.

For our case study with only five months datasets, we contemplated conducting this forecast model with a two ‘datestamps’ period. We attempted to predict the future of February, and since we were working with five months data, we used the “MS” value that corresponded to the start of the month. The outcome of the function was employed as a parameter in our fitted model's predict procedure. This application generated a forecast table with the datetime column “ds”, and three additional columns “yhat”, “yhat lower”, and “yhat upper”. The “yhat” column represented the predicted value of our metric, “yhat lower” column represented the lower limit of our predictions, and “yhat upper” column represented the upper limit of our forecasts. The variance of these Prophet forecast numbers was determined by the Markov chain Monte Carlo (MCMC) algorithm.

Given our forecast data in the function plot components, we retrieved the components that reflected, at the temporal level, how the monthly, weekly, and daily patterns of the time series contributed to the total anticipated values.

In Figure 31, for the blue cluster non-roaming, the monthly chart demonstrated a downward trend over time, the weekday chart demonstrated a rise in the beginning of the week, on Monday and it tended to fall until the end of the week, on Sunday. Hourly, it demonstrated that from 7 a.m. to 12 a.m., there was an upward trend. For the blue roaming cluster, the monthly chart showed as well a downward trend over time, for the weekday, it showed a more sinusoidal behavior that acted from its minimum peak on early Tuesday, and then up until Thursday, starting to go down until mid-Friday, and going up again to its maximum on Sunday. Hourly, it showed its highest values from 10 a.m. to around 9 p.m.

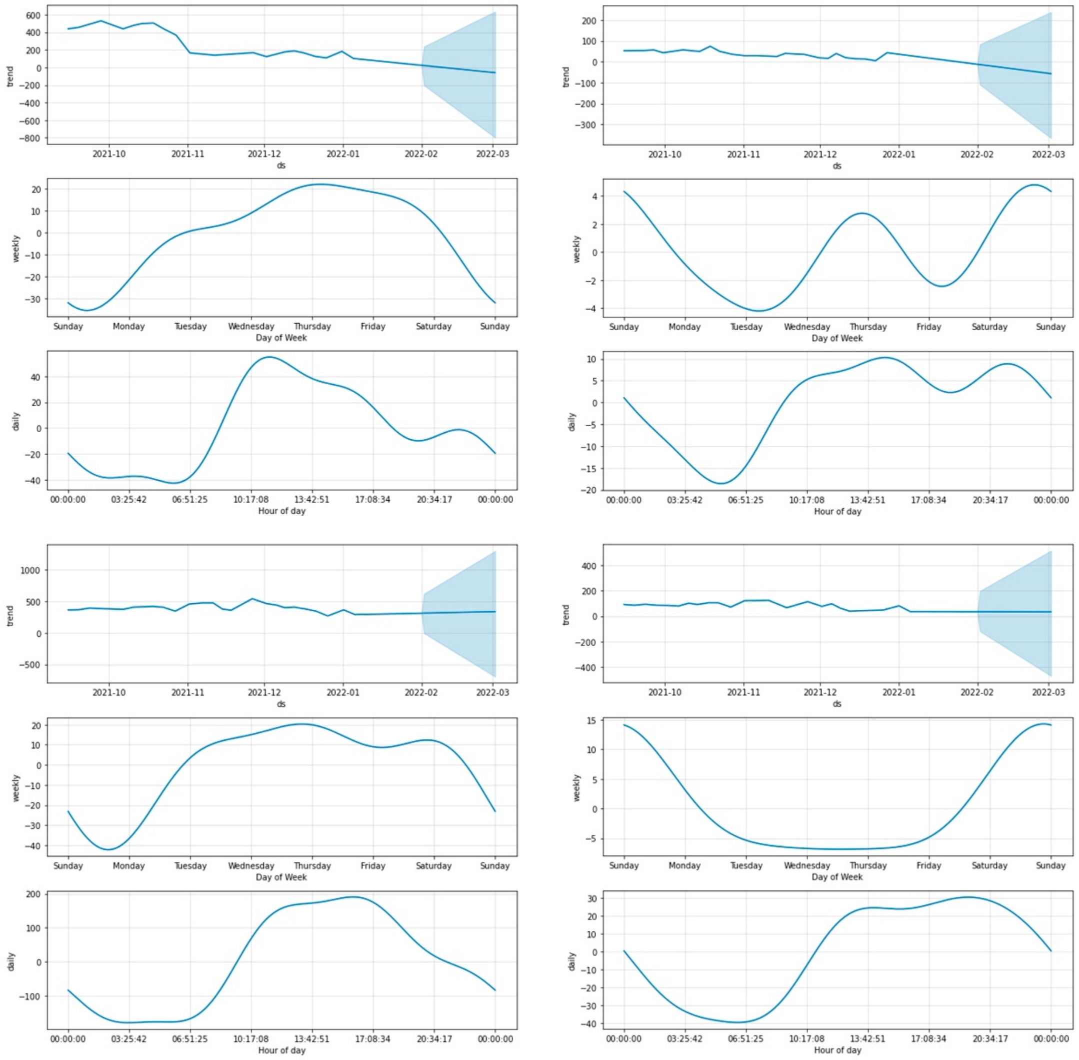


Figure 31 – Blue cluster (first row) and purple cluster (second row) for non-roaming (left) and roaming (right) monthly with February prediction (blue shade), weekly and hourly graphs during all months' dataset

For the purple cluster non-roaming, the monthly model showed a rising trend over time, and weekly a similar result to the blue non-roaming cluster. The hourly model showed a growth, beginning at 7 a.m. and ending at 5 p.m. For the purple cluster roaming, the monthly model showed an increase over time, as in weekly, was quite low during the weekdays, rising on Thursday till Sunday, when it started to fall. Hourly, it started earlier than the non-roaming, reaching a peak at 1 a.m. and falling by 8:30 p.m. The most notable aspects of this analysis were that the blue cluster demonstrated a declining trend, and the purple cluster a positive trend. In contrast to roaming, which tended to be more prevalent on Fridays through Sundays, non-roaming data displayed a considerable presence on weekdays. This mirrors the tourist (roaming) mobility patterns and the resident daily mobility patterns (non-roaming).

Figure 32 showed our four clusters Prophet model in which we observed the values of our time series (the black dots), the forecasted values (blue line), and the forecasted uncertainty ranges (the blue shaded regions). We performed a supplement time series with changepoints, datetime points, where the trend of the time series abruptly changed. These transition points were shown by vertical red dashed lines. Figure 32 showed four changepoints, except for purple cluster roaming with five changepoints, despite the fact that it was programmed to show six, but the changepoint prior scale was set to 0.001, making the trend less flexible, thus it was limited to four monthly seasonality. Then, the period was settled to 30.5 to represent the 30 or 31 days of the month. Seasonality was estimated using the Fourier sum, and the right number for Fourier order in our instance was five. Given the parameterizations described above, the Prophet model of all our cases, showed a downward trend for the month of February.

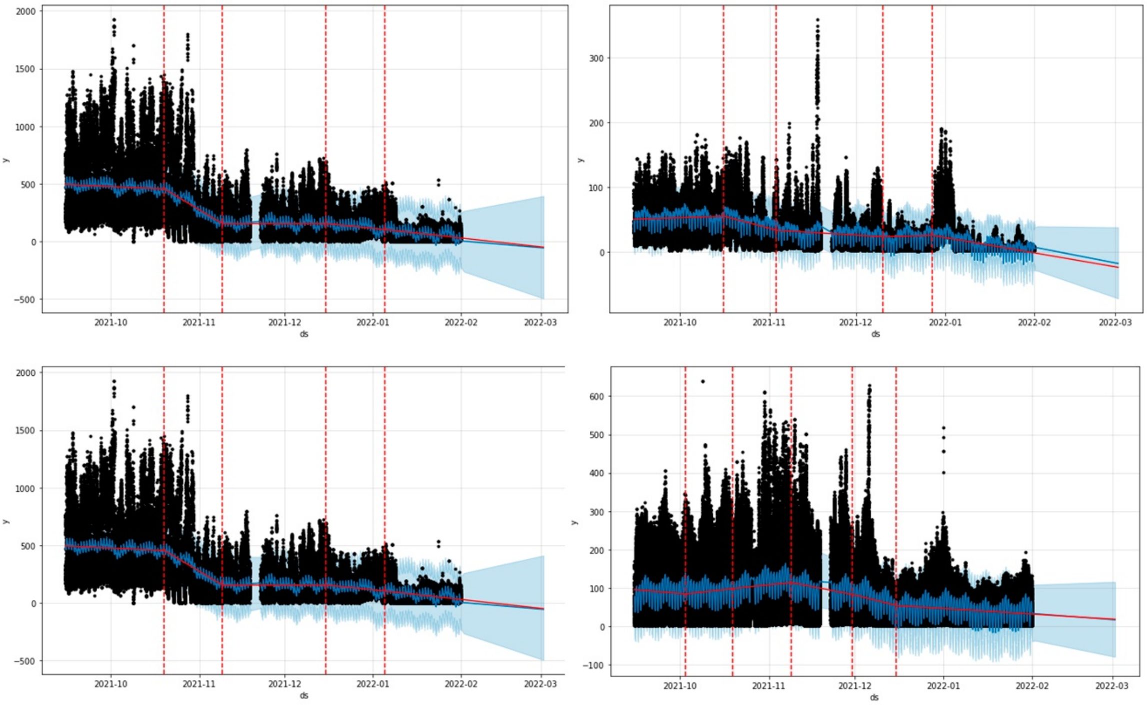


Figure 32 – Prophet plot for blue cluster (first row) and purple cluster (second row) for non-roaming (left) and roaming (right) during all months' dataset with time series (black dots), the forecasted values (blue line), the trend values (red line), the projected uncertainty intervals (blue shaded regions) and the potential changepoints (vertical red lines)

3.5. Evaluation

The results obtained with the implementation of our model allowed us to gather valid information on the various topics addressed in our research question and sub-question that led us to achieve the thesis objectives.

The evaluation of the end-user ensured that the findings were aligned with the proposed research objectives and the business requirements were accurate. The end-user evaluation only contemplated till the DBSCAN model.

We hold a meeting with CML where we presented and discussed the research results and a questionnaire (Table 5) was shared with CML to assess and evaluate the objectives of our results.

*Table 5 – Method assessment questionnaire**

Criteria	Objective statement	Evaluator #1	Evaluator #2
Utility	It can help CML define strategies for urban mobility, in the scope of smartphone users with roaming and non-roaming	FA	LA
Understandability	Provides understandable results	LA	LA
Accessibility	Can be used without training	FA	FA
Level of detail	Provides knowledge to CML in detecting urban mobility patterns with smartphone users' data with roaming and non-roaming	LA	LA
Consistency	Gives consistent results	LA	LA
Robustness	Has enough detail to be used in other Lisbon parishes	FA	FA

**End-user evaluation DBSCAN only*

The development of the questionnaire followed the standards defined by the ISO/IEC TS 33061 [28], used to assess software development processes. The NLPF's four levels used for evaluation:

- Not Achieved (NA) - [0-15%]
- Partially Achieved (PA) -]15-50%]
- Largely Achieved (LA) -]50-85%]
- Fully Achieved (FA) -]85-100%]

Two CML evaluators graded FA in the categories of accessibility and robustness, and LA in the categories of understandability, level of detail and consistency. In the utility category, the evaluation differed between LA and FA rates.

This evaluation suggested that this study helped CML in the strategies for urban mobility, and provided understandable results and knowledge to CML in detecting urban mobility patterns with smartphone users' data (roaming and non-roaming)

The CML evaluators believed that this study in Santa Maria Maior has enough detail to be replicated to other Lisbon's parishes.

Overall, the results were consistent with the goals and specifications put forward for this study, and the outcomes were aligned with the objectives and requirements proposed.

In the CML meeting Engineer João Tremoceiro, director of Lisboa Inteligente, proposed the following actions for future work: analysis of the daily peaks in the parish of Santa Maria Maior, morning, afternoon, and night, and take into account the fact that Santa Maria Maior has extrapolated data, i.e. has a higher volume of activity than the rest of the city of Lisbon; also, Engineer Nuno Ferreira, approached the topic of mobile phone users flow on maritime routes, notably cruises and ferries. This can also be regarded essential for the comprehension of maritime mobility patterns and investigated in future work.

3.6. Deployment

The results were not applied in a real production environment and were compiled in the thesis writing and presented to the CML.

All software development was done in Python on a personal computer with Windows 10 (64bits) operating system, AMD Ryzen 7 5800X 8-Core Processor 3.80GHz, with 32Gb of memory ram. We adopted the python programming language (v3.9.12) [29], compiled with Jupyter Notebooks [30]. The packages used are Numpy [31], Pandas [32], GeoPandas [33], Seaborn [34], Matplotlib [35], Folium [36], datetime [37], contextily [38], Geojson [39], OSMnx [40], Plotly [41], si-prefix [42] and Prophet [43]. The reproducibility of this process can be accomplished by running the Jupyter Notebooks [44].

All the developed software material and data sets are available to use by the CML and for academic research purposes.

4. Conclusions

4.1. Discussion

We addressed our main research question: “what are the mobility patterns of smartphone users in the city of Lisbon, related to points of interest in the city, namely historic places and public transportation?” with the objective to understand the mobility patterns in Lisbon through mobile phone data, in the vicinity of the historic and transportation POIs.

We performed analysis and visualizations of the mobile phone data to identify mobility patterns in Lisbon, supported by the methodologies found in our SLR, such as data mining and visualization. In our data mining approach, we adopted the CRISP-DM methodology [6], [7], using statistical analysis and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [12]–[15].

First, we evaluated which was the Lisbon parish the suited the most for our case study. We conducted an analysis of all Lisbon's parishes generating choropleths for all months with the total number of devices for non-roaming and roaming, followed by choropleths visualizing the number of historic and transportation POIs. Santa Maria Maior stood out for having the highest number of POIs and one of the highest number of devices throughout the months. Following this result, we chose Santa Maria Maior parish to be our case study of our data model.

It was performed week, and hour analysis in section 3.4.1 to understand the mobility patterns of smartphone in Santa Maria Maior. In the weekly view (A 3), for non-roaming, all months showed growth from Monday through Friday, extending to Saturday in October and January, except December, which maintains constant levels throughout the week. Apart from October, all roaming months grew over the week (decreased from Sunday to Wednesday and rose until Saturday). We established a distinct analysis for weekdays and weekends in comparison (A 5). For non-roaming weekdays and weekends followed the same pattern across all months. September and October had more devices between 7 p.m. and 6 a.m., November and December between 4 p.m. and 7 a.m., and January between 5 p.m. and 7 a.m. In December, weekends surpassed weekdays. From 400 devices in October, we saw 450 in November, 550 in December, and under 400 in January. For roaming months, the pattern was similar, except in September, when the average number of weekends roaming devices is higher. Finally, we conducted an analysis of daily hours (A 4) and observed that the day peak for non-roaming devices ranges from 350 in January to 375 in October to 450 in November and December. For roaming data, we saw the lowest activity at 6 a.m. (a little more than 30 devices), with the highest occurring 14 hours later at 8 p.m. (70 devices).

Moreover, we applied DBSCAN to investigate the number of devices related to historical and transportation POIs. The DBSCAN analysis addressed our main research question where standardized clustering POIs revealed interesting relationships with the number of devices. First, we developed four scalar DBSCAN clustering experiments in order to obtain progressively better results. In the first experiment (Figure 26), only historic POIs were used; in the second (Figure 27), only bus stops; in the third (Figure 28), a combination of historic POIs and bus stops; and in the fourth (Figure 29), we included historic, bus stops, metro and train stations POIs. This resulted in DBSCAN models of different combinations of POIs categories that led us to conclude that the optimal number of clusters was eight, of which two clusters (blue cluster and purple cluster), stood out due to their proximity to the 28E tram route, and the Lisbon Cruise Terminal.

This research produced an innovative study, as there were few studies in this field and complemented a published paper [21] that also used Lisbon as a case study but differs on the scope of the type of transport, using public transportation data rather than the shared transportation. Also, looked at the parish of Santa Maria Maior, which was chosen due to the large number of devices for non-roaming and roaming users, as well as the large number of POIs, rather than Beato, Marvila and Parque das Nações. This study also offered significant scientific value by separating the non-roaming and roaming analyses.

Moreover, in the evaluation phase of our DBSCAN model we got positive feedback with categories of accessibility and robustness rated with FA, and LA in the categories of understandability, level of detail and consistency. The CML evaluators believed that this study in Santa Maria Maior had enough detail to be replicated to other Lisbon's parishes.

The results were consistent with the research question put forward for this study, and the outcomes are aligned with the objectives and requirements proposed.

The DBSCAN results were attractive enough throughout development to allow us to move beyond what was established by our SLR and build forecasting algorithm with Prophet implementation.

Following the DBSCAN analysis, we addressed the sub-question: "How can we forecast mobility patterns of smartphone users in Santa Maria Maior?", by applying a forecast model to the purple and blue clusters to determine future mobility patterns.

To accomplish this, we utilized data from all months, referencing the quadrants in which the cluster points were located, therefore separating the analysis from the forecast with non-roaming and roaming data. For this forecasting research, we used the Prophet algorithm based on the analyzed months and performed a forecast only for the month of February, which resulted in a downward projection. This evaluation of this algorithm took into account the data temporal limitation and the anomalous month values resulting from the pandemic restrictions.

Furthermore, our thesis provided answers to our main research question, which concerned the mobility patterns with mobile phone data related to historical and transportation POIs, by applying DBSCAN, as well as, to our sub-question with a prediction algorithm with Prophet resulting in a downward trend of the mobile phone users (non-roaming and roaming) for the month of February.

4.2. 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.

4.3. 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. Future work should also look at the analysis of a full year dataset not compromised with the restrictions of the pandemics. This would allow a more complete and closer to a real scenario as well as provide a better forecast of the model.

References

[1] ‘THE 17 GOALS | Sustainable Development’. https://sdgs.un.org/goals (accessed Aug. 30, 2022).

[2] ‘70/303. Modalities for the United Nations Conference to Support the Implementation of Sustainable Development Goal 14: Conserve and sustainably use the oceans, seas and marine resources for sustainable development’.

[3] ‘Plataforma de Gestão Inteligente de Lisboa - Lisboa Inteligente’. https://lisboainteligente.cm-lisboa.pt/lxi-iniciativas/plataforma-de-gestao-inteligente-de-lisboa/ (accessed Oct. 21, 2022).

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

[5] ‘FCT — About FCT’. https://www.fct.pt/fct.phtml.en (accessed Oct. 21, 2022).

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

[7] 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.

[8] 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.

[9] ‘VOSviewer - Visualizing scientific landscapes’. https://www.vosviewer.com/ (accessed Aug. 30, 2022).

[10] 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.

[11] ‘Mendeley’. https://www.mendeley.com/search/ (accessed Aug. 30, 2022).

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] ‘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).

[24] ‘Evolução da mobilidade na cidade de Lisboa face às medidas de desconfinamento – LxDataLab’. https://lisboainteligente.cm-lisboa.pt/lxdatalab/desafios/evolucao-da-mobilidade-na-cidade-de-lisboa-face-as-medidas-de-desconfinamento/ (accessed Oct. 26, 2022).

[25] ‘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

[26] ‘Folium — Folium 0.12.1 documentation’. http://python-visualization.github.io/folium/ (accessed Sep. 09, 2022).

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

[28] ‘ISO - ISO/IEC TS 33061:2021 - Information technology — Process assessment — Process assessment model for software life cycle processes’. https://www.iso.org/standard/80362.html (accessed Oct. 04, 2022).

[29] 'Python Release Python 3.9.12 | Python.org'. https://www.python.org/downloads/release/python-3912/ (accessed Oct. 22, 2022).

[30] 'Project Jupyter | Home'. https://jupyter.org/ (accessed Oct. 22, 2022).

[31] 'NumPy'. https://numpy.org/ (accessed Oct. 22, 2022).

[32] 'pandas - Python Data Analysis Library'. https://pandas.pydata.org/ (accessed Oct. 22, 2022).

[33] 'GeoPandas 0.11.0 — GeoPandas 0.11.0+0.g1977b50.dirty documentation'. https://geopandas.org/en/stable/ (accessed Oct. 22, 2022).

[34] 'seaborn: statistical data visualization — seaborn 0.12.1 documentation'. https://seaborn.pydata.org/ (accessed Oct. 22, 2022).

[35] 'Matplotlib — Visualization with Python'. https://matplotlib.org/ (accessed Oct. 22, 2022).

[36] 'Folium — Folium 0.12.1 documentation'. http://python-visualization.github.io/folium/ (accessed Oct. 22, 2022).

[37] 'datetime — Basic date and time types — Python 3.10.8 documentation'. https://docs.python.org/3/library/datetime.html (accessed Oct. 22, 2022).

[38] 'contextily: context geo tiles in Python — contextily 1.1.0 documentation'. https://contextily.readthedocs.io/en/latest/ (accessed Oct. 22, 2022).

[39] 'geojson · PyPI'. https://pypi.org/project/geojson/ (accessed Oct. 22, 2022).

[40] 'OSMnx 1.2.2 — OSMnx 1.2.2 documentation'. https://osmnx.readthedocs.io/en/stable/ (accessed Oct. 22, 2022).

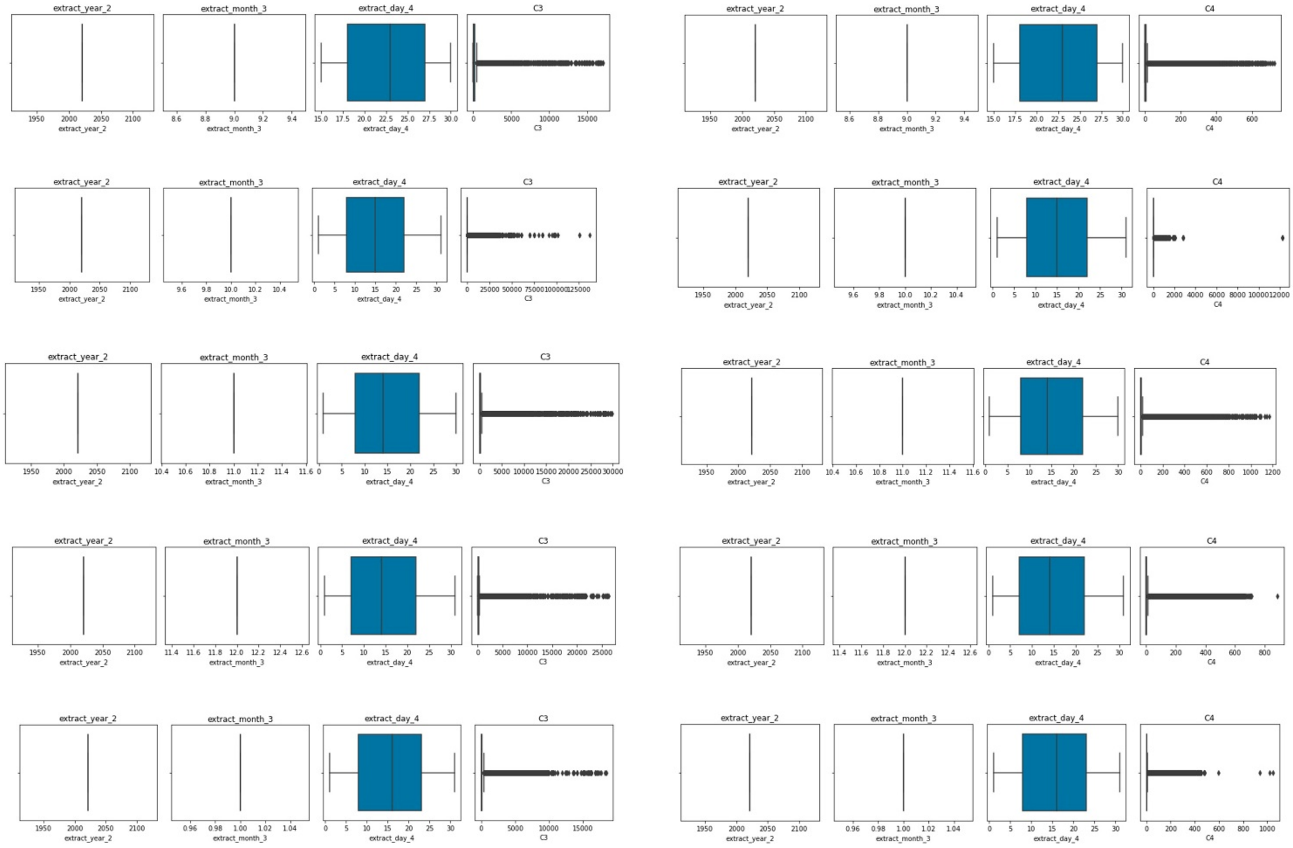
[41] 'Getting started with plotly in Python'. https://plotly.com/python/getting-started/ (accessed Oct. 22, 2022).

[42] 'si-prefix · PyPI'. https://pypi.org/project/si-prefix/ (accessed Oct. 22, 2022).

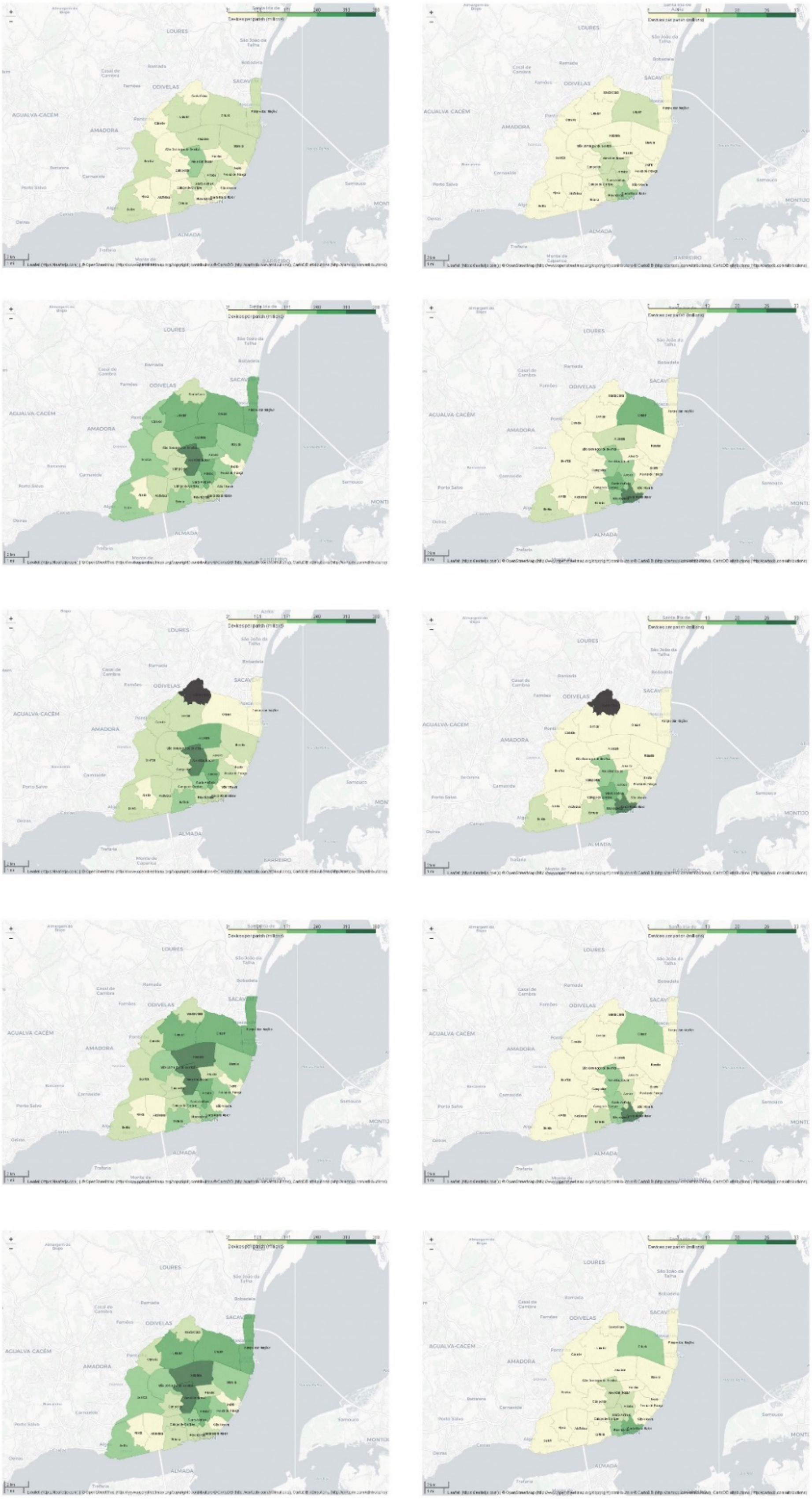
[43] 'Quick Start | Prophet'. https://facebook.github.io/prophet/docs/quick_start.html (accessed Oct. 22, 2022).

[44] 'Project Jupyter | Home'. https://jupyter.org/ (accessed Oct. 22, 2022).

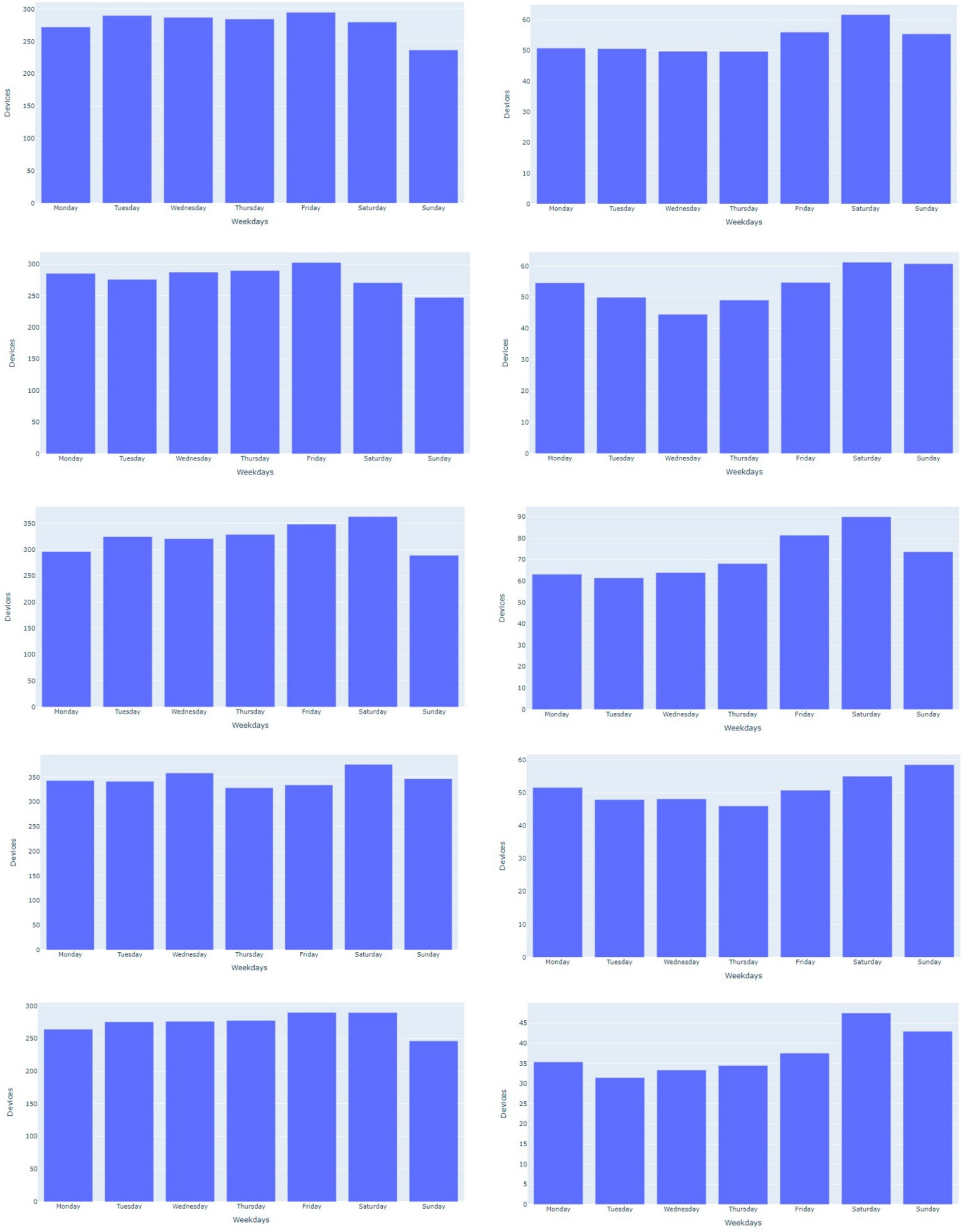
Appendix



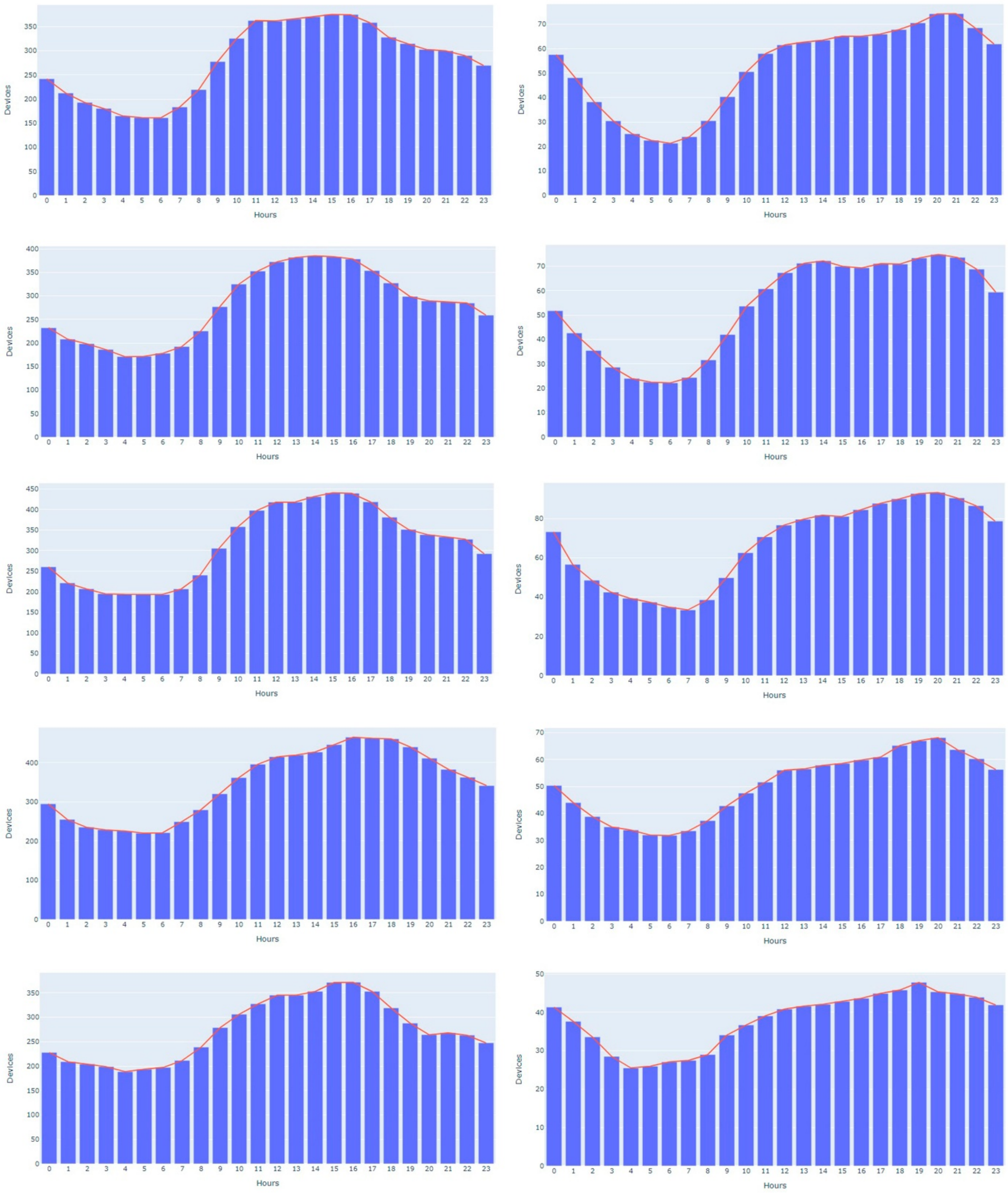
A 1 – Boxplots for outlier detection in all-monthly datasets non-roaming (left) and roaming (right)



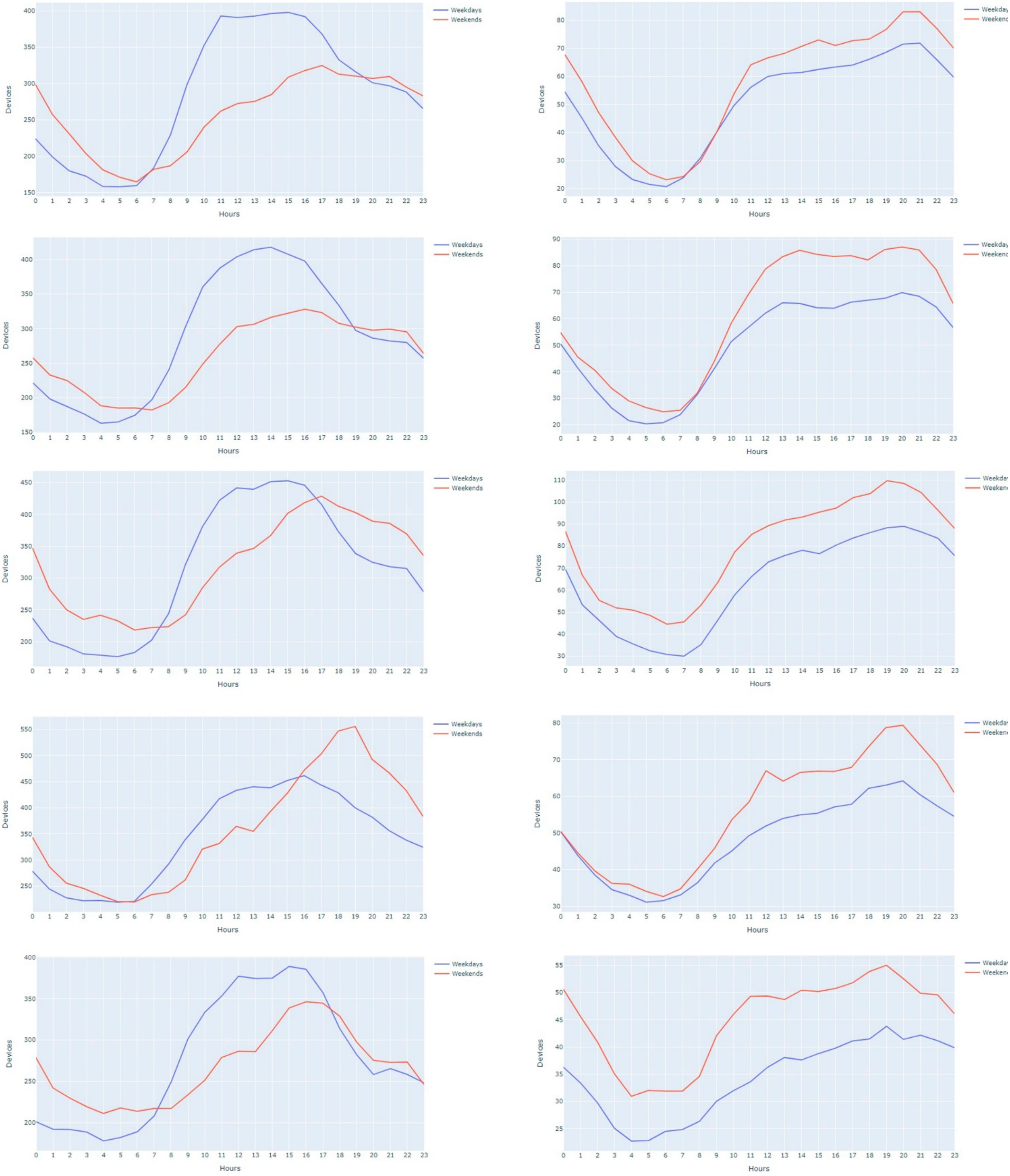
A 2 – Devices (millions) per parish in all-monthly datasets non-roaming (left) and roaming (right)



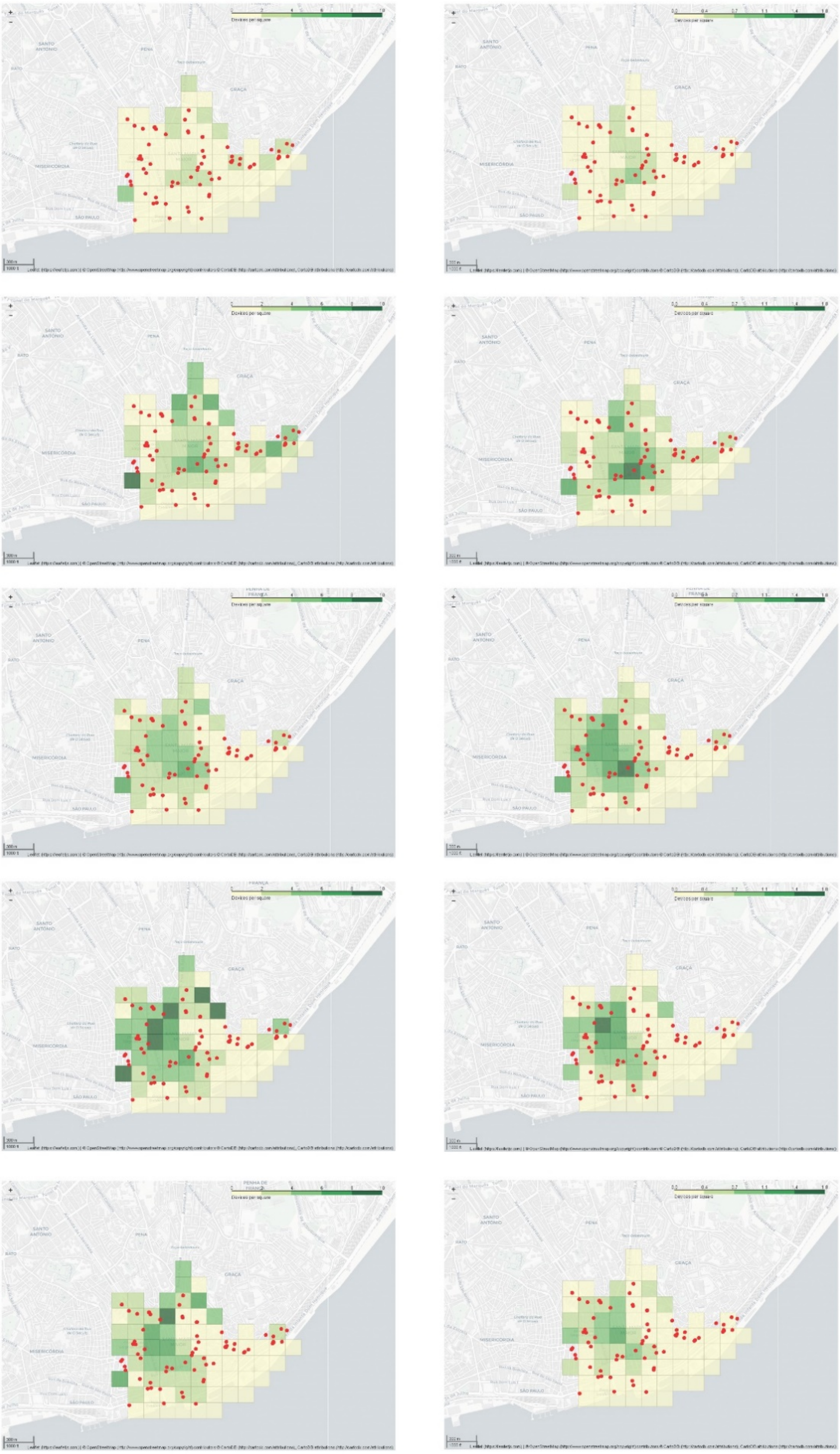
A 3 – Santa Maria Maior's average weekdays devices in all-monthly datasets non-roaming (left) and roaming (right)



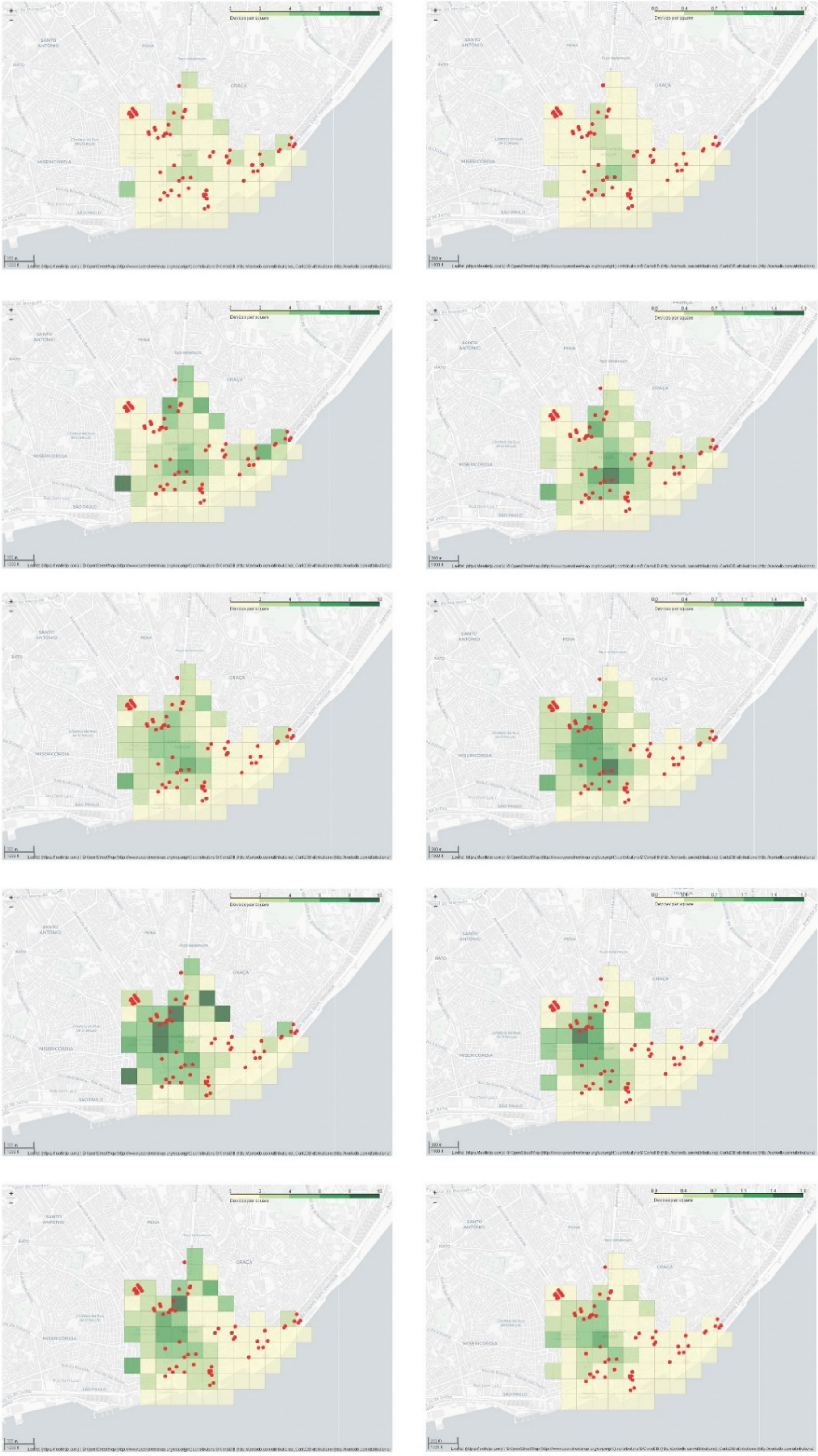
A 4 – Santa Maria Maior's average daily hours' devices in all-monthly datasets non-roaming (left) and roaming (right)



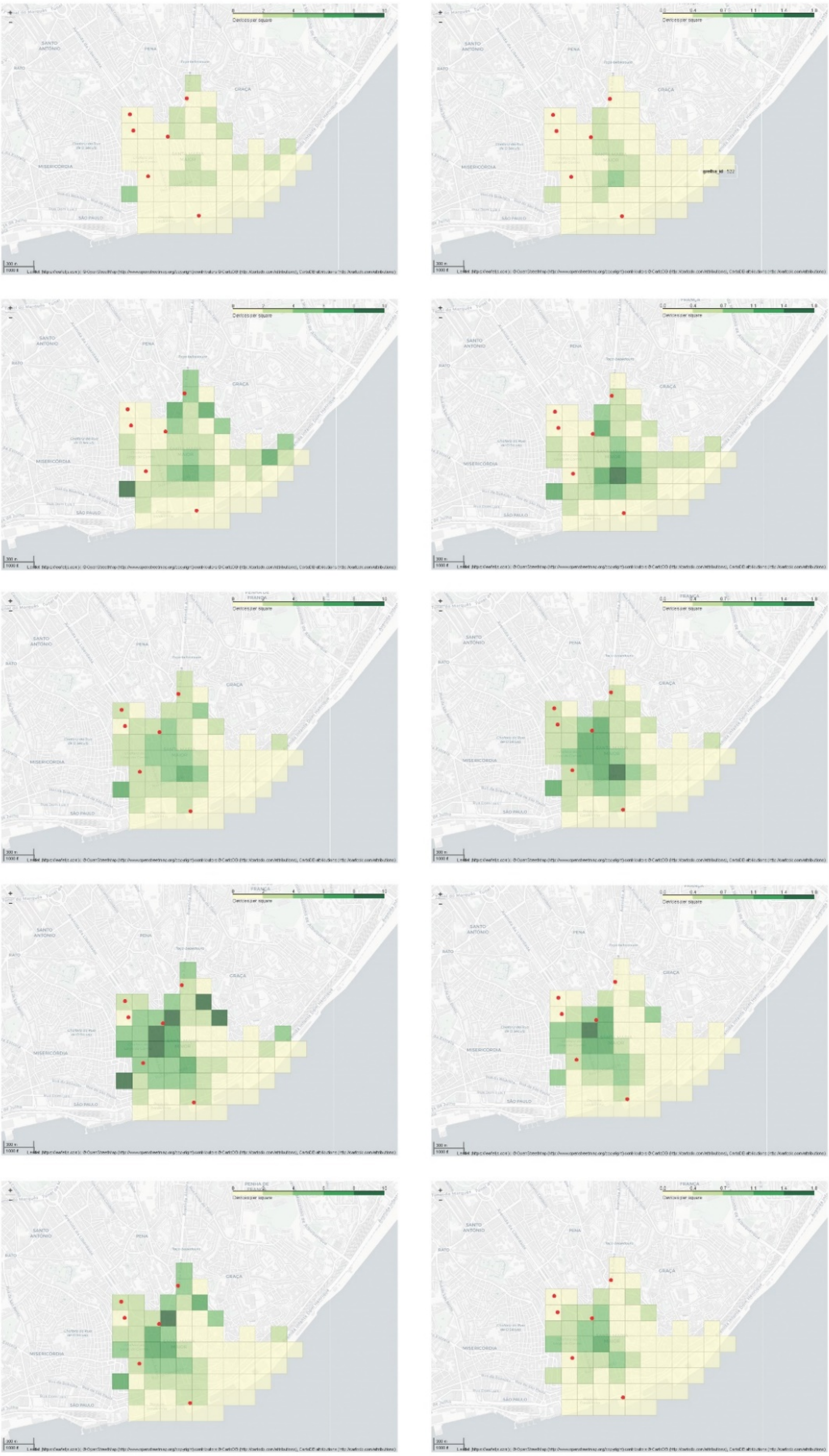
A 5 – Santa Maria Maior's average weekdays and weekends' devices in all-monthly datasets non-roaming (left) and roaming (right)



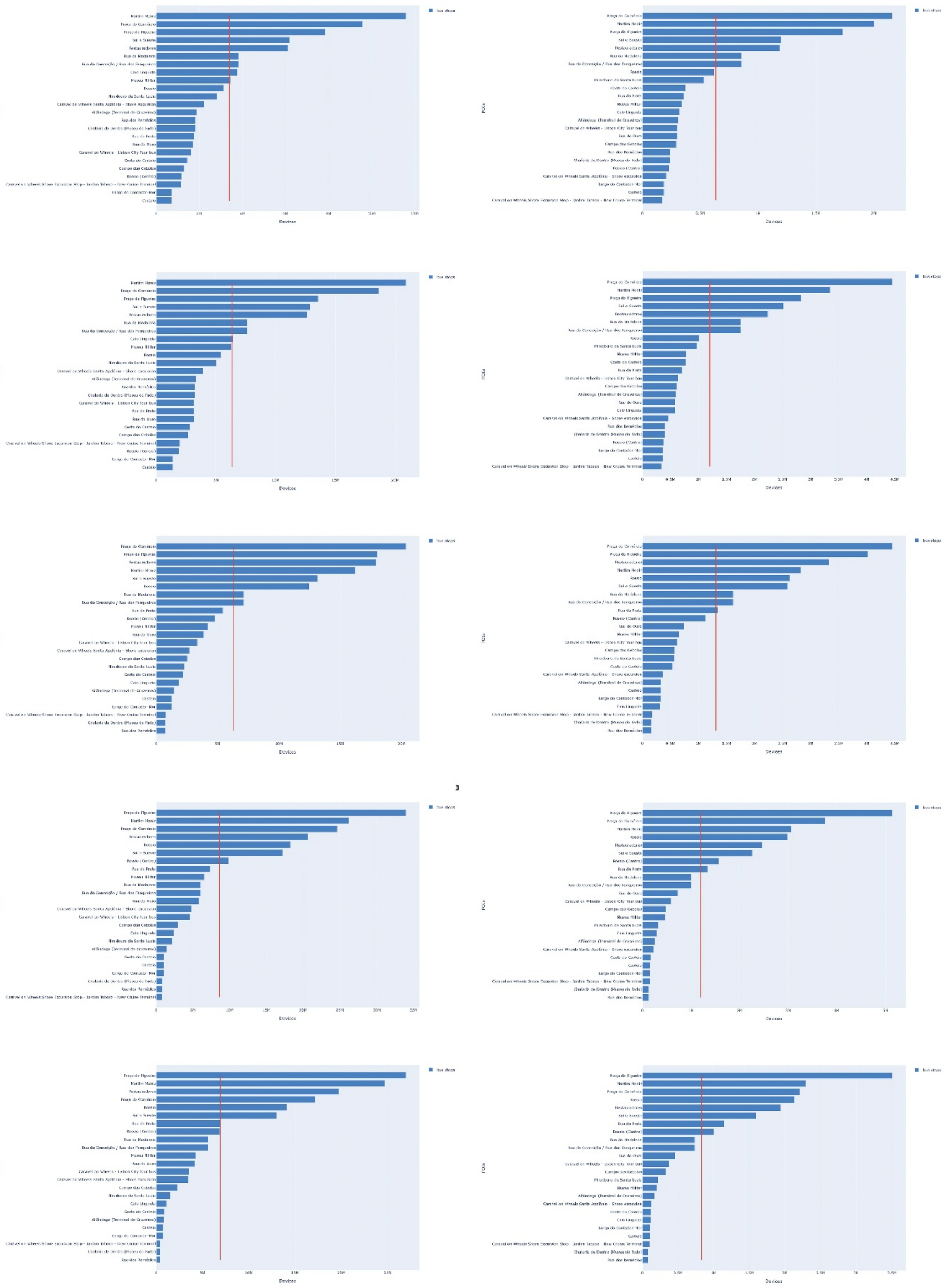
A 6 – Historic POIs influence area (red dots) in Santa Maria Maior's smartphone data quadrants (millions) in all-monthly datasets non-roaming (left) and roaming (right)



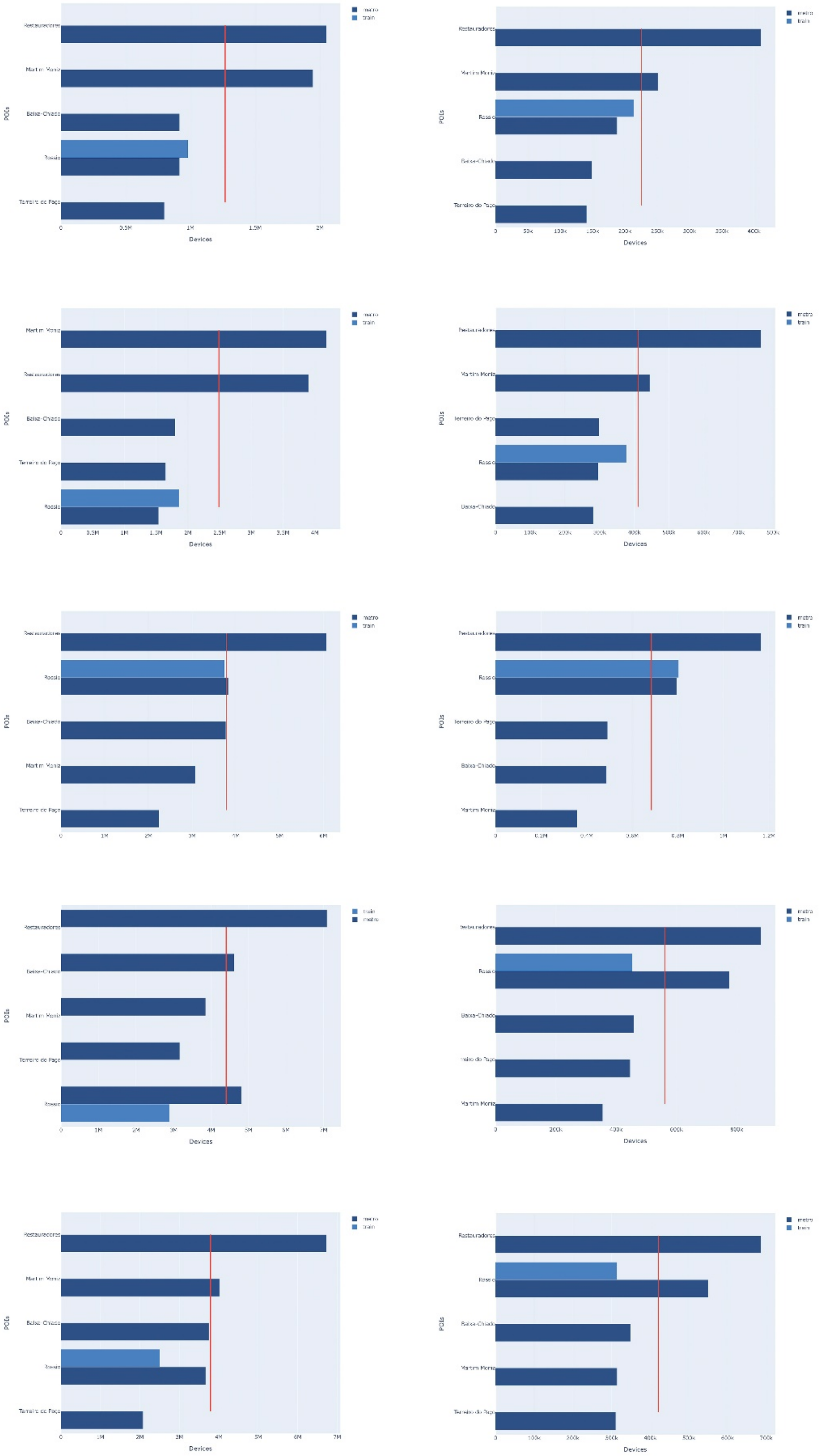
A 7 – Bus stops POIs influence area (red dots) in Santa Maria Maior's smartphone data quadrants (millions) in all-monthly datasets non-roaming (left) and roaming (right)



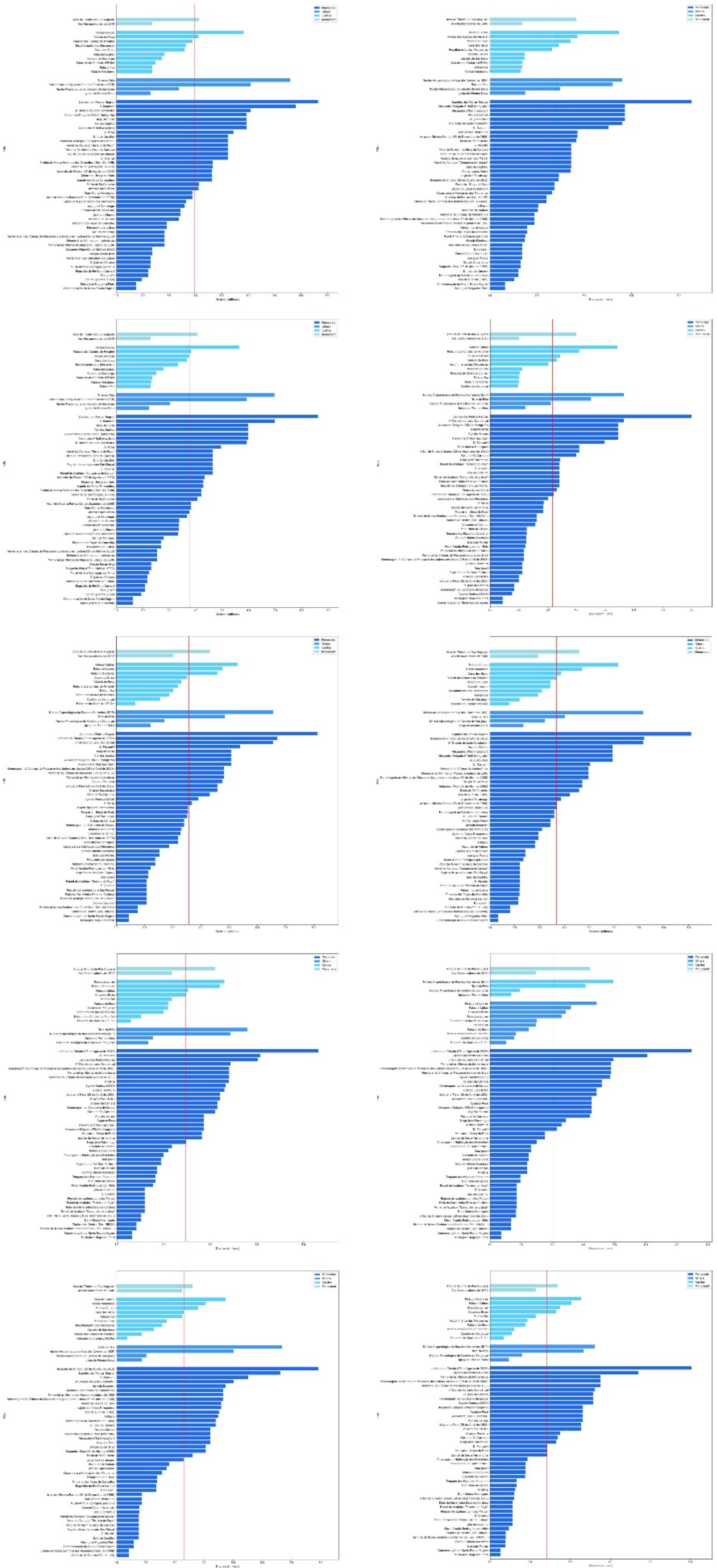
A 8 – Train and metro POIs influence area (red dots) in Santa Maria Maior's smartphone data quadrants (millions) in all-monthly datasets non-roaming (left) and roaming (right)



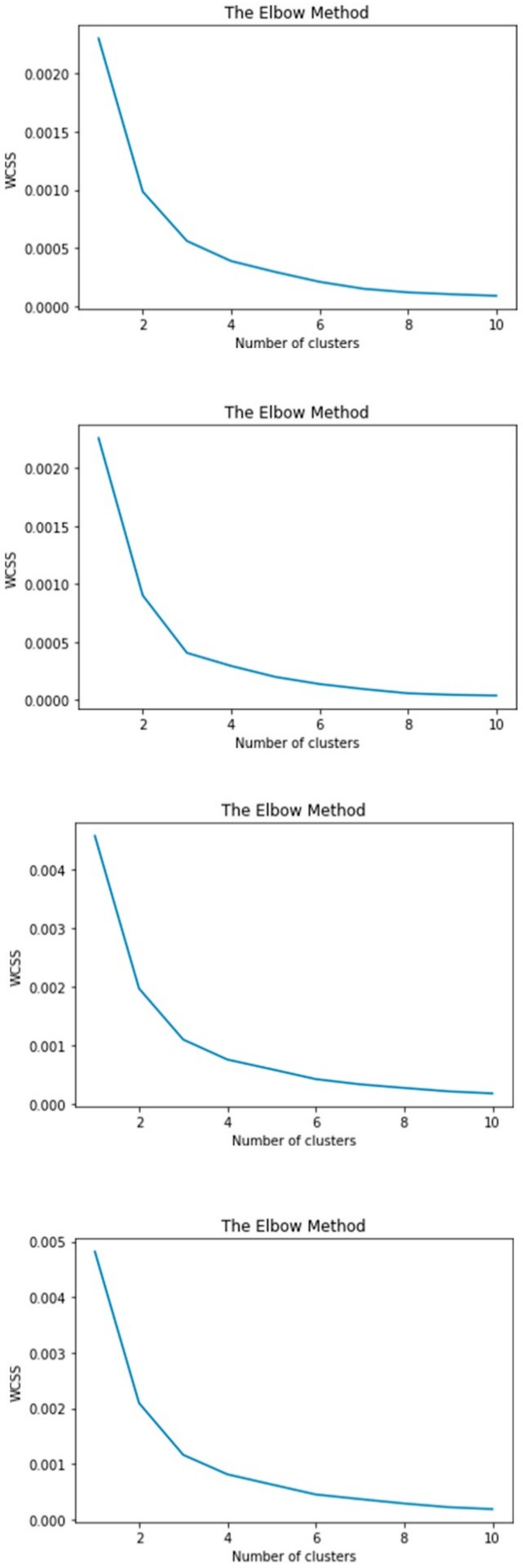
A 9 – Santa Maria Maior bus stops POIs with devices' average (red line) in all-monthly datasets non-roaming (left) and roaming (right)



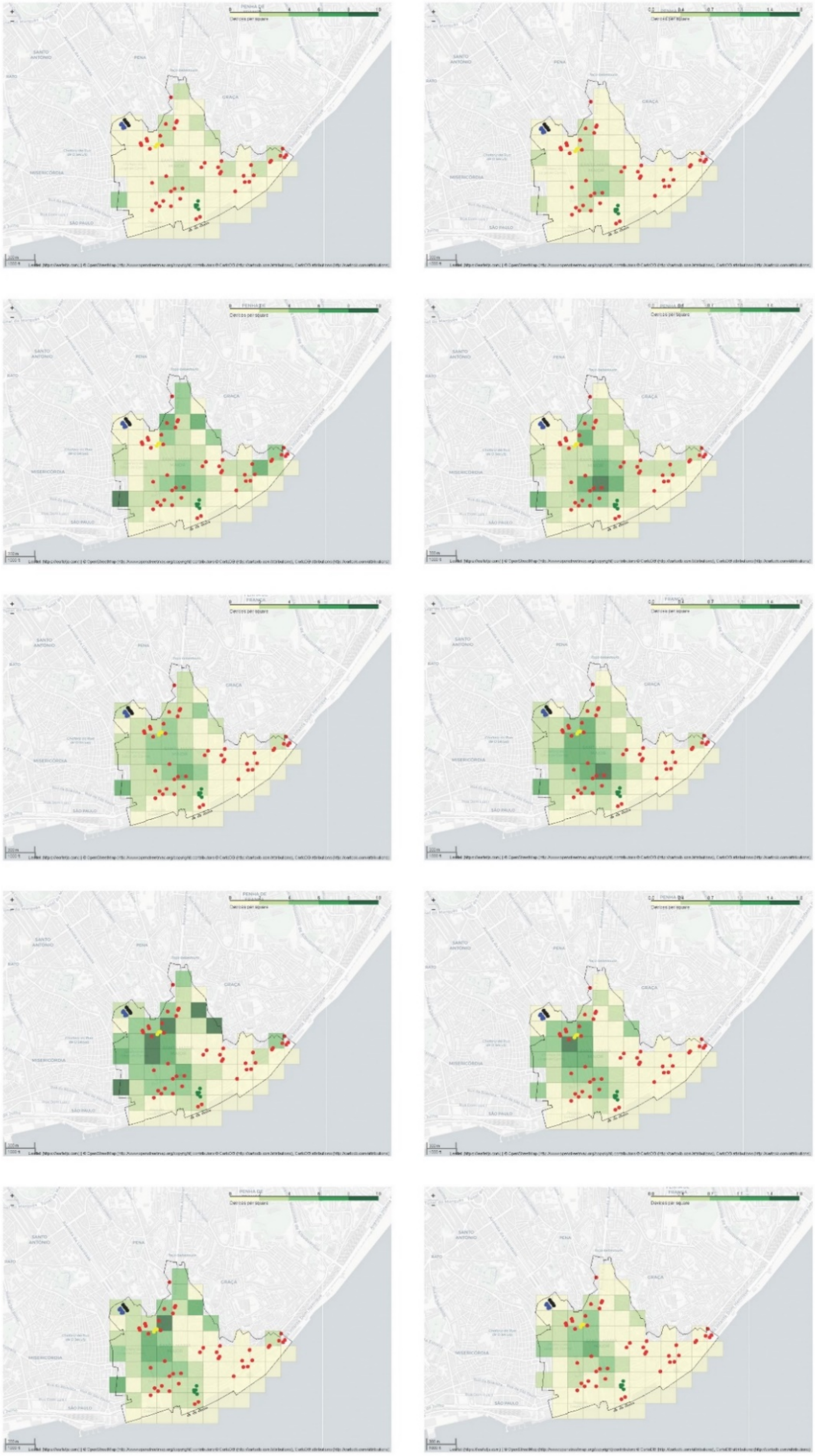
A 10 – Santa Maria Maior train and metro POIs with devices' average (red line) in all-monthly datasets non-roaming (left) and roaming (right)



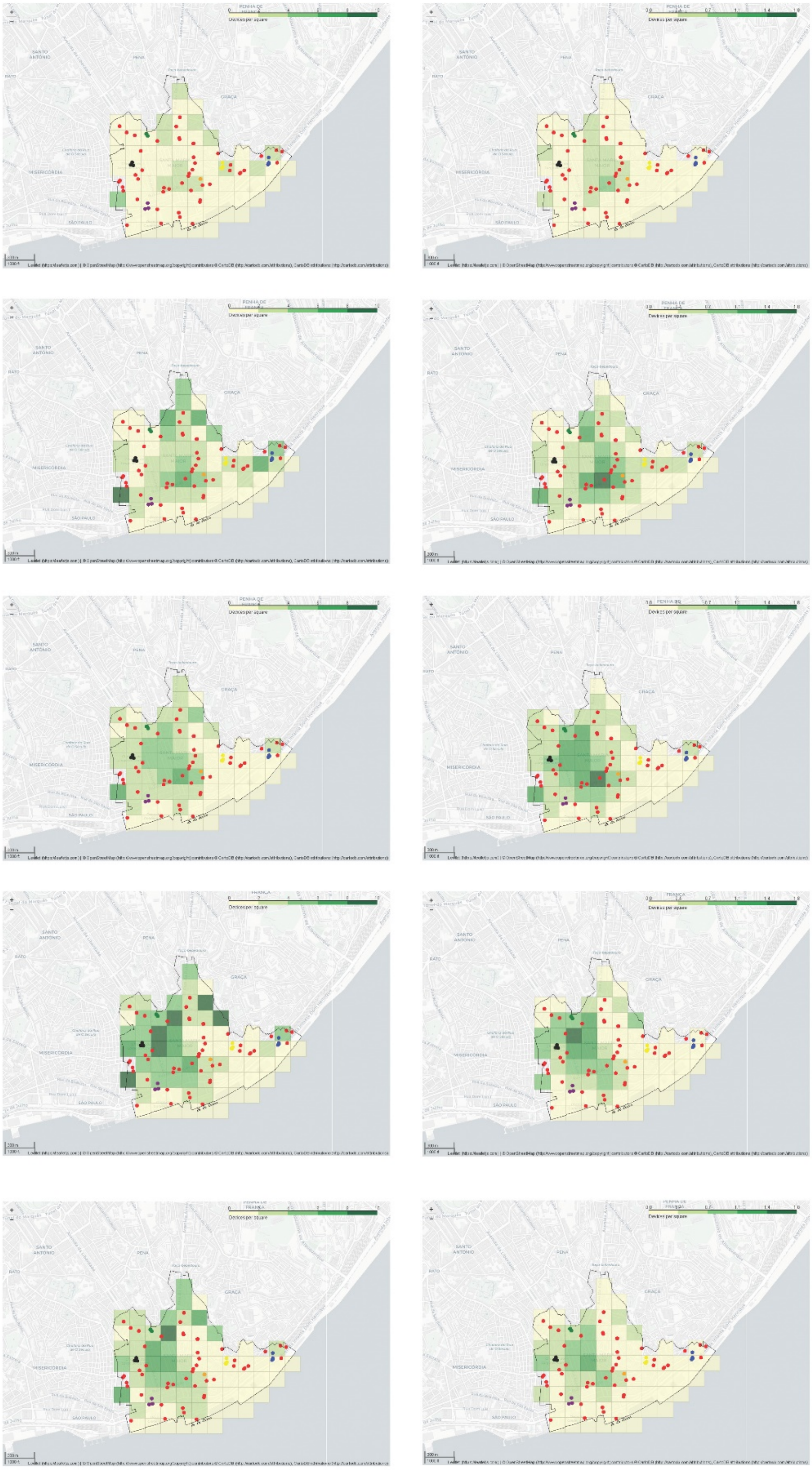
A 11 – DBSCAN bus stops devices (millions) in Santa Maria Maior’s smartphone data quadrants in all-monthly datasets non-roaming (left) and roaming (right)



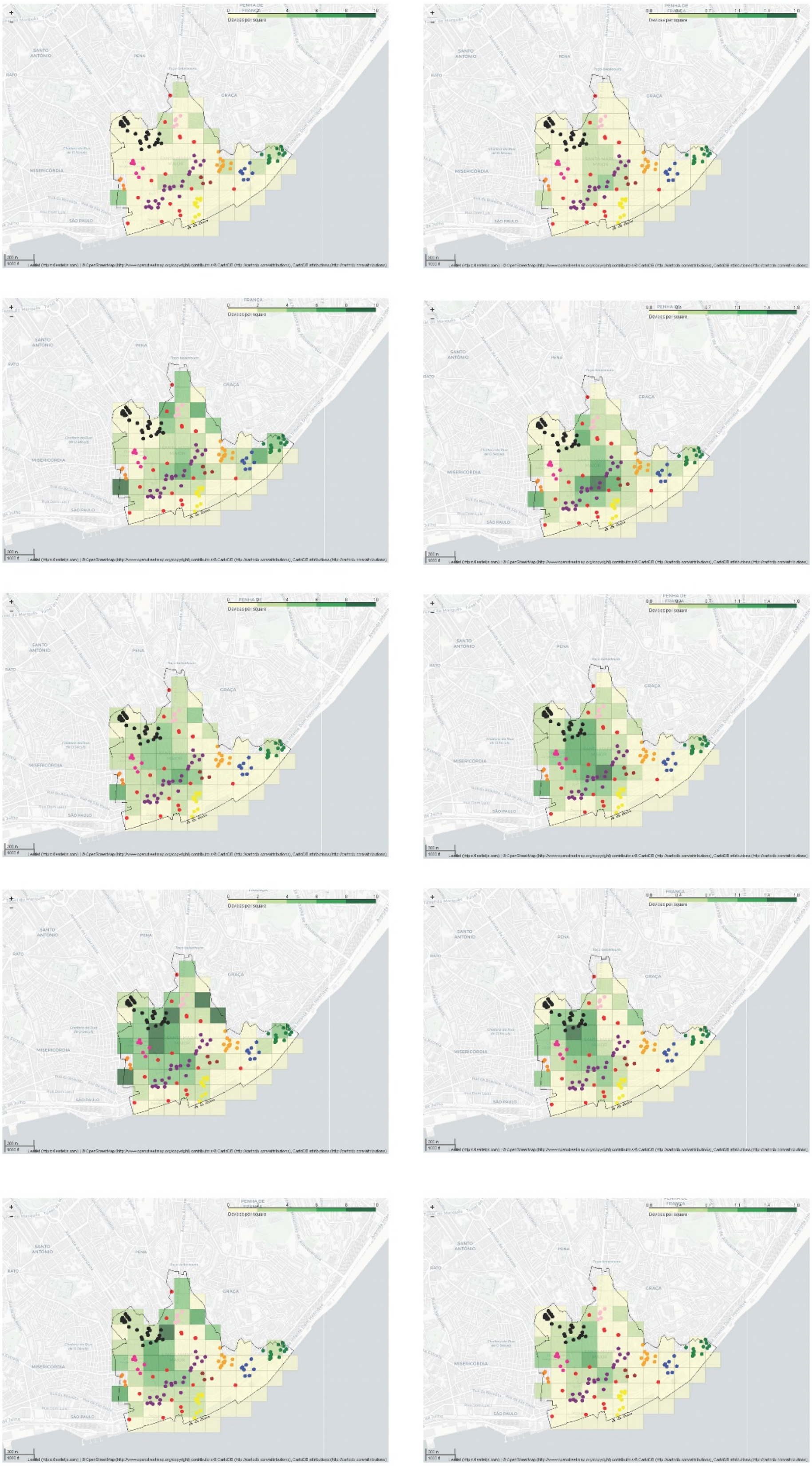
A 12 – Elbow methods respectively for historic, bus, historic and bus, and all POIs



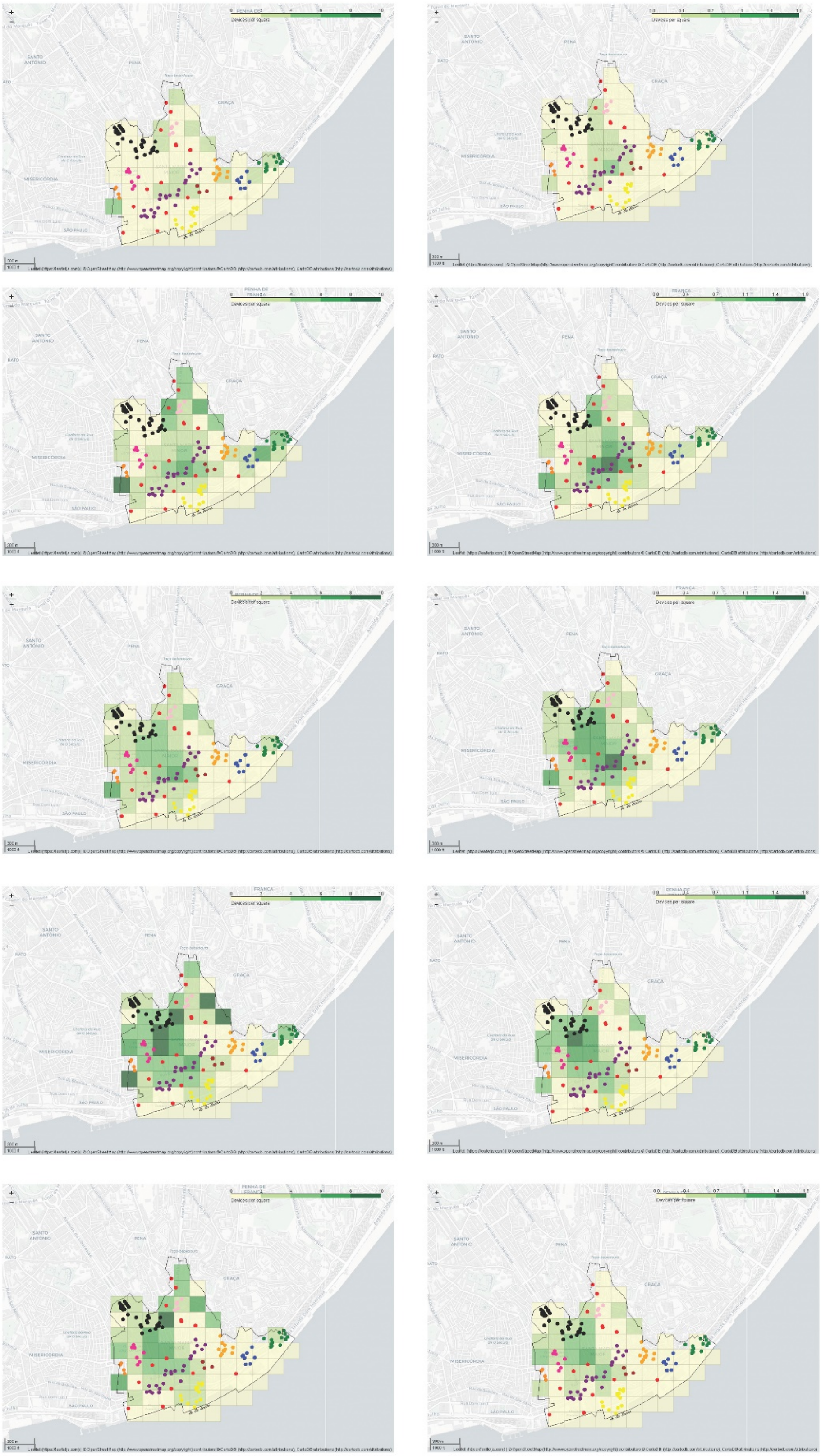
A 13 – DBSCAN bus stops devices (millions) in Santa Maria Maior's smartphone data quadrants (millions) in all-monthly datasets non-roaming (left) and roaming (right)



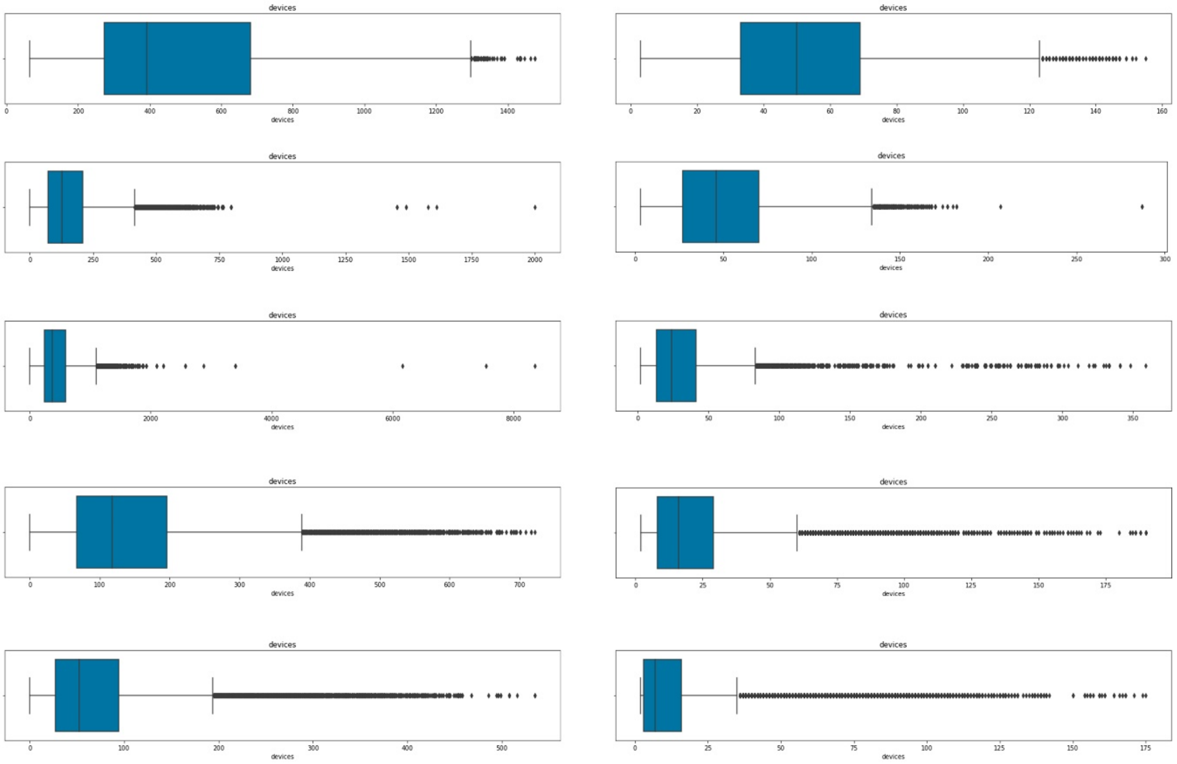
A 14 – DBSCAN historic devices (millions) in Santa Maria Maior's smartphone data quadrants (millions) in all-monthly datasets non-roaming (left) and roaming (right))



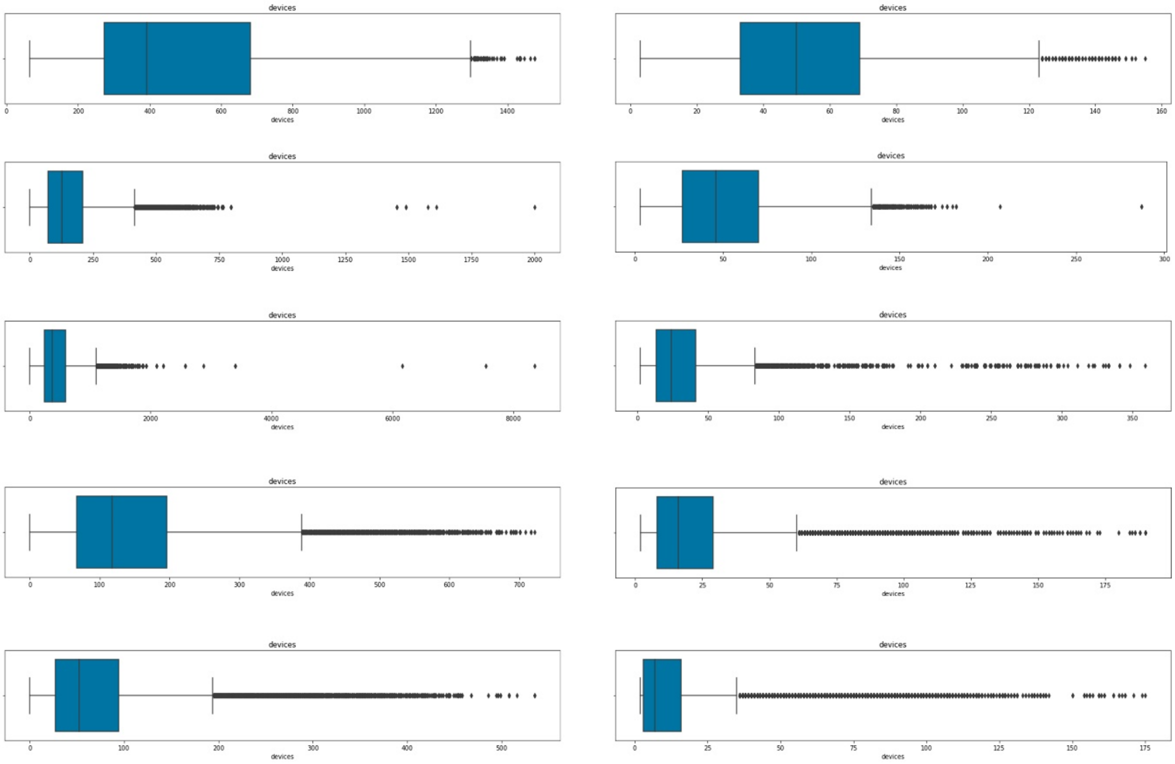
A 15 – DBSCAN bus and historic devices (millions) in Santa Maria Maior's smartphone data quadrants (millions) in all-monthly datasets non-roaming (left) and roaming (right)



A 16 – DBSCAN historic, bus stops and railway stations devices (millions) in Santa Maria Maior's smartphone data quadrants (millions) in all-monthly datasets non-roaming (left) and roaming (right)



A 17 – Blue cluster boxplots for outlier analysis in all-monthly datasets non-roaming (left) and roaming (right)



A 18 – Purple cluster boxplots for outlier analysis in all-monthly datasets non-roaming (left) and roaming (right)