NOVA IMS Information Management School

MEGI

Master Degree Program in

Statistics and Information Management

Urban Mobility Pattern Using Mobile Phone Geolocation Data

A Smart Mobility Case Study of Oeiras

Andre Felipe Saraiva de Melo

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Statistics and Information Management

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

Urban Mobility Pattern Using Mobile Phone Geolocation Data

A Smart Mobility Case Study of Oeiras

by

Andre Felipe Saraiva de Melo

Master Thesis Work presented as partial requirement for obtaining the Master's degree in Statistics and Information Management, with a specialization in Data Analytics

Supervised by

Supervisor: André Barriguinha, PhD, NOVA Information Management School

July, 2025

STATEMENT OF INTEGRITY

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

Lisbon, July 2025

Andre Felipe Saraiva de Melo

DEDICATION

I dedicate this work to my parents, who worked tirelessly from Sunday to Sunday and, even in the face of hardship, always found ways to provide for my education and to share their worldview, where anything is possible through effort and love.

I also dedicate it to Lucianna, who encouraged me and helped me overcome the barriers I once thought impossible to break.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my academic advisor, Professor André Barriguinha, for his guidance, encouragement, and invaluable feedback throughout the development of this thesis. His expertise and support were fundamental at every stage of the research process.

I am also grateful to NOVA Information Management School (NOVA IMS) for providing an excellent academic environment, resources, and the opportunity to pursue this work as part of my master's studies.

Finally, I extend my appreciation to all those who, directly or indirectly, contributed to this work.

ABSTRACT

This study investigates urban mobility patterns in Oeiras, Portugal, by applying data science techniques to anonymized mobile phone geolocation data from 2024 to enhance the municipality's Sustainable Urban Mobility Plan (SUMP). Framed within a Smart Mobility and CRISP-DM methodology, this research analyzes user activity and origin-destination datasets aggregated by census statistical sections. The methodology involved preprocessing the data, followed by Principal Component Analysis (PCA) for dimensionality reduction and K-means clustering to segment the municipality based on mobility behaviors. The PCA successfully reduced the data's complexity into three components that explain 82.4% of the variance, representing the intensity of local activity, connectivity with Lisbon, and connections to the broader Lisbon Metropolitan Area. The subsequent K-means analysis identified four distinct clusters: areas of high local activity with limited external connections (Cluster 1); low-activity baseline areas (Cluster 2); zones with high local activity and strong connections to Lisbon (Cluster 3); and transit areas with significant connectivity to the wider metropolitan region but low internal activity (Cluster 4). A comparison with the 2019 Oeiras mobility survey confirms these patterns while adding significant spatial granularity. The findings provide actionable, evidence-based insights for optimizing public transport, promoting active mobility, and informing land-use planning, demonstrating the value of integrating big data analytics with traditional methods for sustainable urban development.

KEYWORDS

Urban Mobility; Smart Cities; Smart Mobility; Mobile Phone Data; Cluster Analysis

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AMU	Moderately Urban Areas
APR	Predominantly Rural Areas
APU	Predominantly Urban Areas
CRISP-DM	CRoss-Industry Standard Process for Data Mining
CSV	Comma-Separated Values
GDPR	General Data Protection Regulation
GPKG	GeoPackage
ICTs	Information and Communication Technologies
INE	Instituto Nacional de Estatística (National Institute of Statistics)
IoT	Internet of Things
LMA	Lisbon Metropolitan Area
NUTS	Nomenclature of Territorial Units for Statistics
OD	Origin-Destination
PC	Principal Component
PCA	Principal Component Analysis
SDG	Sustainable Development Goals
SSE	Sum of Squared Errors
SUMP	Sustainable Urban Mobility Plan

1. INTRODUCTION

1.1. CONTEXT

In recent decades, the world has experienced a significant shift toward urbanization. According to the United Nations, the proportion of the global population residing in cities increased from 25% in 1950 to approximately 50% in 2020 and is projected to rise gradually to 58% by 2070 (Oginga Martins & Sharifi, 2022).

While urbanization is closely linked to economic progress within the context of globalization, it also presents adverse environmental effects and a decline in overall quality of life (Dociu & Dunarintu, 2012). One of the most critical areas affected is mobility, as urban development is associated with increased motorization (Fornalchyk & Hilevych, 2023), as well as longer commuting distances, a higher incidence of traffic accidents, increased congestion, and worsening air pollution (Gao & Zhu, 2022).

Given the adverse impacts of urbanization, policymakers are actively seeking strategies to mitigate them. To address these challenges, the European Commission introduced the Sustainable Urban Mobility Plan (SUMP) (Rupprecht et al., 2019), a strategic framework for sustainable urban mobility planning. It aims to enhance accessibility and quality of life in cities by balancing economic development, social inclusion, and environmental protection.

SUMPs prioritize public transportation, active mobility (including walking and cycling), and innovative solutions such as shared transportation and vehicle electrification. Furthermore, they emphasize the involvement of stakeholders and citizens in the planning process, ensuring that solutions are tailored to local needs. The overarching goal is to reduce congestion, lower pollutant emissions, and improve the efficiency of urban travel, ultimately making cities more sustainable and livable (Rupprecht et al., 2019).

In this context, Oeiras, a city in Portugal's Lisbon metropolitan area with a population of 175,677 in 2023, developed its Sustainable Urban Mobility and Accessibility Plan (Município de Oeiras, 2022). The project is based on an initial diagnosis that considers the territorial structure and land use patterns, the existing transportation and parking networks, the 2021 Portuguese Census (Instituto Nacional de Estatística, 2022), and the mobility trends and dynamics identified through a public survey conducted in 2019. These insights informed the development of intervention strategies and goals aimed at increasing public transportation and active mobility usage, reducing traffic accidents, and lowering greenhouse gas emissions (Município de Oeiras, 2022).

The term 'Smart Mobility' refers to a new approach for addressing urban challenges through the use of advanced technologies, such as the Internet of Things (IoT) and Information and Communication Technologies (ICTs) (Benevolo et al., 2016). More specifically, Smart Mobility aims to enhance traffic management and gain a deeper understanding of mobility patterns in cities, thereby optimizing infrastructure planning and decision-making (Allam & Sharifi, 2022).

1.2. RESEARCH OBJECTIVES

This study investigates how Smart Mobility approaches, leveraging data science techniques applied to anonymized mobile phone geolocation data, can generate novel insights to inform and enhance Oeiras' mobility planning initiatives. By analyzing mobility patterns through computational methods, this research seeks to provide evidence-based recommendations that complement traditional survey-based approaches to urban planning.

The research addresses a critical gap in the existing literature by applying advanced analytical techniques to Oeiras' specific urban context. Although similar methodologies have been successfully implemented in other cities, including Lisbon, the unique characteristics of Oeiras warrant a dedicated investigation. This analysis directly supports the municipality's evolving mobility objectives and provides actionable insights for sustainable urban development.

To achieve these goals, this study establishes three interconnected research objectives that build upon each other to provide a comprehensive understanding of urban mobility dynamics in Oeiras:

1. Analyze urban mobility patterns by identifying regions with similar behavior throughout the week. This includes exploring the spatial and temporal dimensions of geolocation sections to uncover mobility trends across different city zones.
2. Implement clustering methodologies to classify urban areas based on shared mobility characteristics. This involves evaluating different clustering strategies, defining meaningful cluster interpretations, and assessing their effectiveness in capturing urban dynamics.
3. Analyze the results in comparison with the findings of the previous mobility survey to identify trends and disparities. Assess their alignment with the strategies proposed in Oeiras' Sustainable Urban Mobility Plan (SUMP) and generate actionable insights to enhance urban planning and optimize the distribution of people within the city.

These research objectives collectively aim to contribute to both academic knowledge and practical urban planning applications. From a methodological perspective, the study demonstrates the application of big data analytics to suburban mobility contexts, provides validated frameworks for integrating traditional and computational mobility analysis methods, and establishes best practices for anonymized telecommunications data utilization in urban planning. In terms of practical applications, the research delivers actionable insights to support Oeiras' sustainable mobility initiatives, informs infrastructure investment decisions through data-driven mobility pattern identification, supports evidence-based policy development for urban transportation planning, and enhances understanding of suburban mobility dynamics in the Portuguese context. The outcomes of this study will serve as a

foundation for ongoing mobility monitoring in Oeiras while contributing to the broader discourse on Smart Cities and data-driven urban planning methodologies.

1.3. THESIS STRUCTURE

This research is structured as follows:

- **Chapter 2**: A literature review exploring previous studies on urban dynamics and tracking methods, comparing various analytical approaches.
- **Chapter 3**: A description of the dataset used in the study, including its sources and characteristics.
- **Chapter 4**: A detailed explanation of the methodology, outlining the steps taken to process the data and generate results.
- **Chapter 5**: Presentation of the study's findings, supported by visual representations such as graphs and maps illustrating the urban structure.
- **Chapter 6**: Discussion of the results, emphasizing their implications for urban planning and policy-making.
- **Chapter 7**: Conclusion summarizing key insights and proposing directions for future research.

2. LITERATURE REVIEW

2.1. SMART CITIES AND SMART MOBILITY

The concept of smart cities lacks a universally agreed-upon definition, yet common themes emerge across various studies. Broadly, smart cities utilize information and communication technologies (ICTs) to enhance urban services, improve quality of life, and increase operational efficiency while minimizing resource consumption and environmental impact (Mohanty et al., 2016). Key components include intelligent infrastructure, transportation, energy management, healthcare, and digital technology (Mohanty et al., 2016). The Internet of Things (IoT) and big data analytics play a crucial role in making cities more adaptive and responsive (Mohanty et al., 2016).

A more comprehensive definition by Batra & Chhabra (2023) emphasizes the integration of innovation and ICTs to address present and future social, environmental, and economic challenges. Similarly, the IEEE IoT Initiative's Smart Cities Working Group defines smart cities as urban environments designed to enhance citizen well-being, improve service delivery, and strengthen citizen-government interaction (Hammons & Myers, 2019).

Within this broader framework, smart mobility emerges as a key pillar in addressing urban transportation challenges through advanced digital technologies (Mitieka et al., 2023). By integrating IoT, big data analytics, and artificial intelligence, smart mobility solutions aim to improve transportation efficiency, accessibility, and sustainability (Mitieka et al., 2023; Porru et al., 2020). Core components include intelligent transport systems, open data initiatives, and citizen engagement (Bıyık et al., 2021).

Despite its potential, several challenges hinder the widespread adoption of smart mobility solutions. These include integrating new technologies with existing transport networks, standardizing metrics for optimal route detection, and navigating the complexities of investment planning (Porru et al., 2020; Goumiri et al., 2023). Additionally, current research often prioritizes technological advancements while overlooking social and economic dimensions, underscoring the need for more multidisciplinary approaches (Mitieka et al., 2023).

While smart mobility can be applied in both urban and rural settings, implementation varies based on population density and infrastructure availability (Porru et al., 2020). Thus, as cities evolve into smarter ecosystems, mobility plays an increasingly vital role in shaping sustainable and efficient urban environments.

2.2. MOBILITY PATTERNS

Urban mobility patterns are essential for understanding city dynamics and development. Recent studies have leveraged various data sources to analyze human movement in urban environments. Yuan et al. (2012) used mobile phone data to extract dynamic mobility

patterns, employing Dynamic Time Warping to measure similarities between time series. Noulas et al. (2021) analyzed Foursquare data across multiple cities, identifying a universal law of human mobility based on rank-distance rather than physical distance. Their findings suggest that variations in movement patterns stem primarily from the spatial distribution of places within urban areas. Similarly, Hasan et al. (2013) utilized smart subway fare card data to model urban mobility, proposing a simple model based on place popularity to predict visited locations. This model effectively replicated observed travel behaviors, including trip frequency and the exploration of new places. Collectively, these studies highlight the potential of big data in understanding and modeling urban mobility, offering valuable insights for researchers and policymakers.

Mobile phone data, in particular, has proven to be a powerful resource for studying human mobility, with applications in urban planning and transportation management. Researchers have developed various techniques to extract meaningful insights from this data. Some approaches focus on predicting individual movements using Dynamic Bayesian Networks (Dash et al., 2016)**,** while others identify key locations, such as home and work, based on phone usage frequency and duration (Demissie et al., 2019). Clustering techniques have been applied to categorize users into distinct mobility profiles, uncovering weekly patterns and territorial dynamics (Thuillier et al., 2018). Advanced methods, such as Latent Dirichlet Allocation and Affinity Propagation, have been used to extract mobility topics and patterns, addressing challenges related to data sparsity and localization noise (Yang et al., 2019). Despite its limitations in accuracy and detail compared to GPS data, mobile phone data remains a valuable tool for analyzing human mobility (Dash et al., 2016; Demissie et al., 2019).

3. DATASET DESCRIPTION

The dataset used in this study was provided by a mobile network operator in Portugal. It was collected anonymously at regular intervals over the 366 days of the year 2024, primarily covering the municipality of Oeiras and extending to other areas within the Lisbon Metropolitan Area (LMA). The data were geographically aggregated by statistical sections, defined for administrative and operational purposes by the 2021 Census of Portugal (Instituto Nacional de Estatística, 2022).

The dataset comprises two Comma-Separated Values (CSV) files: one containing an origin-destination matrix that records user mobility during specific daily time intervals, and another reporting the number of active users per statistical section captured through periodic snapshots. In addition, a GeoPackage (GPKG) file, based on the 2021 Census, provides spatial features such as the geographic boundaries of statistical sections, represented through multipolygon geometries, as well as contextual sociodemographic attributes.

It's important to note that these datasets are provided at different temporal granularities, hourly for user activity and bi-hourly for movement data. The process of temporal harmonization to ensure consistency for analytical purposes is detailed in the Methodology section.

3.1. CENSUS STATISTICAL SECTIONS

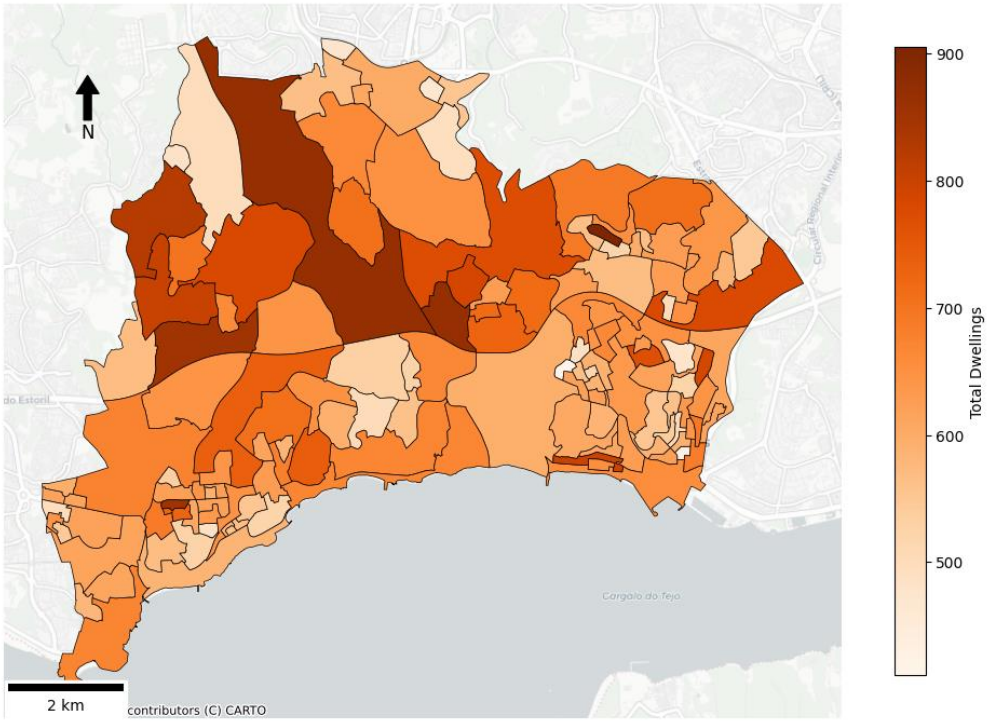

Figure 3-1 – Number of dwellings per statistical section in the municipality of Oeiras

According to the 2021 Census of Portugal, a statistical section is a territorial unit composed of a continuous area located entirely within a single parish, typically comprising between 550 and 650 residential dwellings. The average number of dwellings depends on the parish's degree of urbanization: approximately 550 in Predominantly Rural Areas (APR), 600 in Moderately Urban Areas (AMU), and 650 in Predominantly Urban Areas (APU) (INE, 2022). Deviations from these values may occur due to constraints related to geographic continuity and housing distribution, provided they remain within predefined limits for each classification.

The municipality of Oeiras includes 139 statistical sections distributed across its five parishes. Although the sectioning aimed to standardize the number of dwellings, variations persist, with a mean of 624.7 dwellings and a standard deviation of 91.6 (Figure 3-1). Due to differences in settlement density, the surface area of the sections varies substantially, with an average of 120,009 m² and a standard deviation of 497,912 m².

In addition to dwelling counts, the 2021 Census provides a wide range of contextual variables, including data aligned with the Nomenclature of Territorial Units for Statistics (NUTS), the hierarchical classification system used by the European Union. It also includes information on the number and characteristics of conventional buildings and dwellings, family nuclei, and individual residents. These variables vary across statistical sections and are used to normalize mobility and location data, supporting comparative spatial analysis. Each section is assigned a unique nine-digit code that facilitates integration with other project datasets. The 39 census features available are described in Table 3-1.

Table 3-1 – Description of 2021 Portugal Census Dataset

Feature	Description
NUTS1 DSG	NUTS1 region designation - Major Region
NUTS2 DSG	NUTS2 region designation - Region
NUTS3 DSG	NUTS3 region designation - Sub-region
MUNICIPIO DSG	Municipality designation
FREGUESIA DSG	Parish designation
SECCAO	Statistical section
N_EDIFICIOS_CLASSICOS	Number of classic buildings
N_EDIFICIOS_CLASS_CONST_1_OU_2_ALOJ	Number of classic buildings with 1 or 2 dwellings
N_EDIFICIOS_CLASS_CONST_3_OU_MAIS_ALOJAMENTOS	Number of classic buildings with 3 or more dwellings
N_EDIFICIOS_EXCLUSIV_RESID	Number of exclusively residential buildings
N_EDIFICIOS_1_OU_2_PISOS	Number of buildings with 1 or 2 floors
N_EDIFICIOS_3_OU_MAIS_PISOS	Number of buildings with 3 or more floors
N_EDIFICIOS_CONSTR_ANTES_1945	Number of buildings constructed before 1945

N_EDIFICIOS_CONSTR_1946_1980	Number of buildings constructed 1946-1980
N_EDIFICIOS_CONSTR_1981_2000	Number of buildings constructed 1981-2000
N_EDIFICIOS_CONSTR_2001_2010	Number of buildings constructed 2001-2010
N_EDIFICIOS_CONSTR_2011_2021	Number of buildings constructed 2011-2021
N_EDIFICIOS_COM_NECESSIDADES_REPARACAO	Number of buildings needing repairs
N_ALOJAMENTOS_TOTAL	Total number of dwellings
N_ALOJAMENTOS_FAMILIARES	Number of family dwellings
N_ALOJAMENTOS_FAM_CLASS_RHABITUAL	Number of family dwellings classified as usual residence
N_ALOJAMENTOS_FAM_CLASS_VAGOS_OU_RESID_SECUNDARIA	Number of family dwellings vacant or secondary residence
N_RHABITUAL_ACESSIVEL_CADEIRAS_RODAS	Number of usual residences accessible to wheelchairs
N_RHABITUAL_COM_ESTACIONAMENTO	Number of usual residences with parking
N_RHABITUAL_PROP_OCUP	Number of usual residences owner-occupied
N_RHABITUAL_ARRENDADOS	Number of usual residences rented
N_AGREGADOS_DOMESTICOS_PRIVADOS	Number of private domestic aggregates
N_ADP_1_OU_2_PESSOAS	Number of domestic aggregates with 1 or 2 people
N_ADP_3_OU_MAIS_PESSOAS	Number of domestic aggregates with 3 or more people
N_NUCLEOS_FAMILIARES	Number of family nuclei
N_NUCLEOS_FAMILIARES_COM_FILHOS_TENDO_O_MAIS_NOVO_MENOS_DE_25	Number of family nuclei with children, youngest under 25
N_INDIVIDUOS	Number of individuals
N_INDIVIDUOS_H	Number of male individuals
N_INDIVIDUOS_M	Number of female individuals
N_INDIVIDUOS_0_14	Number of individuals aged 0-14
N_INDIVIDUOS_15_24	Number of individuals aged 15-24
N_INDIVIDUOS_25_64	Number of individuals aged 25-64
N_INDIVIDUOS_65_OU_MAIS	Number of individuals aged 65 or more
geometry	Geographic coordinates of the statistical area

3.2. USER ACTIVITY DATA

The User Activity CSV dataset reports the hourly count of active mobile devices within each statistical section in Oeiras over the course of a day. The dataset includes one column identifying the date of data collection and another indicating the corresponding statistical section. The remaining columns represent each of the 24 hours, containing the respective number of active users, as detailed in Table 3-2. In compliance with GDPR regulations, values equal to or less than 5 have been anonymized and replaced with zero to safeguard individual privacy.

Table 3-2 – Description of Fields in the Hourly Active User Dataset

Feature	Description
datakey	Reference date for the data collection
cod_secc	Code identifying the statistical section
val_00h	Number of active users in the section recorded at 00:00
val_01h	Number of active users in the section recorded at 01:00
val_02h	Number of active users in the section recorded at 02:00
…	…
val_21h	Number of active users in the section recorded at 21:00
val_22h	Number of active users in the section recorded at 22:00
val_23h	Number of active users in the section recorded at 23:00

By aggregating the data daily and calculating the average of the total number of active users per hour, an overall daily mean of 459,155 users was observed, with a standard deviation of 46,910. Additionally, clear seasonal patterns are evident at both the monthly and weekly levels, as illustrated in Figure 3-2.

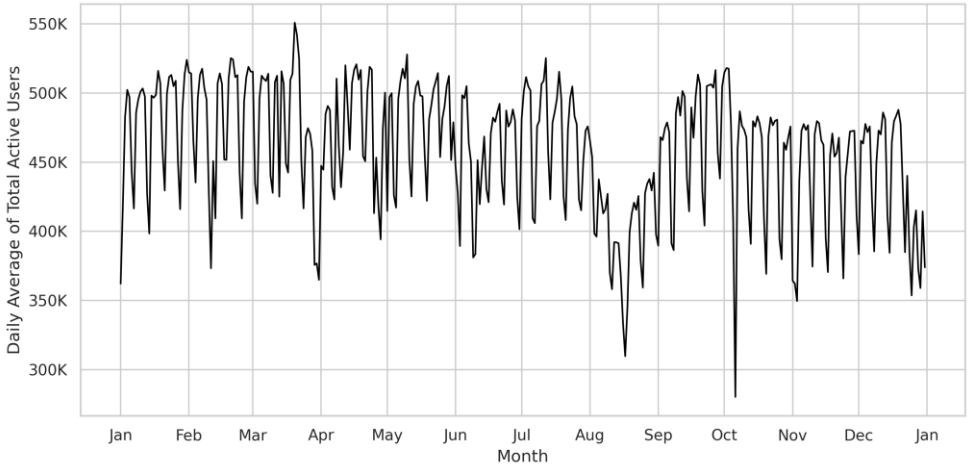

Figure 3-2 – Daily average total number of active users in Oeiras (2024)

Figure 3-3 presents boxplot visualizations that offer a clearer representation of monthly seasonality. A noticeable decline in user activity is observed in June and August, which aligns with holiday periods and summer vacations in Portugal. Additionally, a downward trend in activity is evident toward the final months of the year.

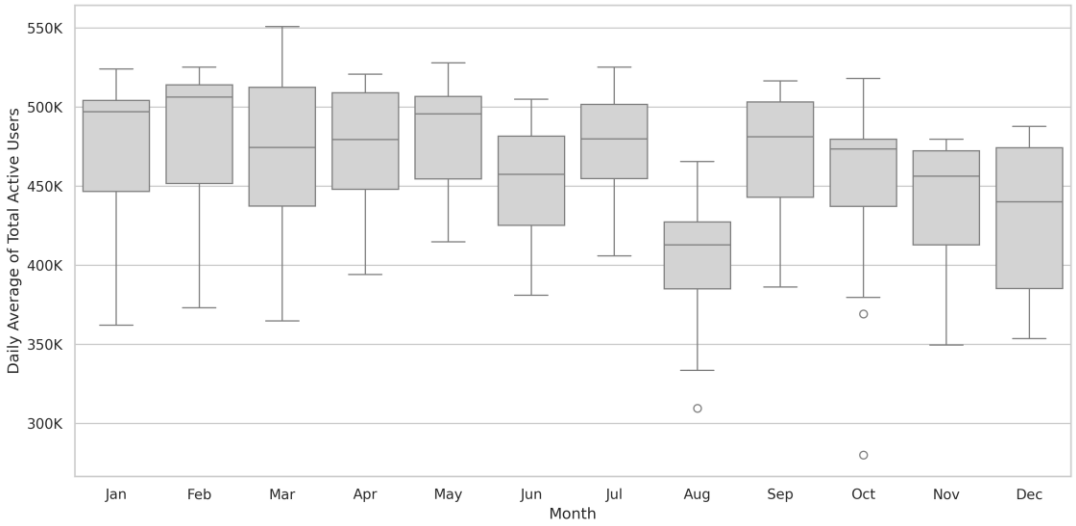

Figure 3-3 – Boxplot of Daily Average Active Users per Month in Oeiras (2024)

Figure 3-4 illustrates the weekly seasonality by presenting the average number of active users per hour for each day of the week. The data reveals a marked distinction between weekdays and weekends, with weekdays, particularly Thursday, showing the highest median levels of user activity, while weekends, especially Sunday, exhibit significantly lower values. Additionally, the presence of lower outliers is observed, corresponding to holiday periods and days in August, reflecting reduced user activity typically associated with vacation times in Portugal.

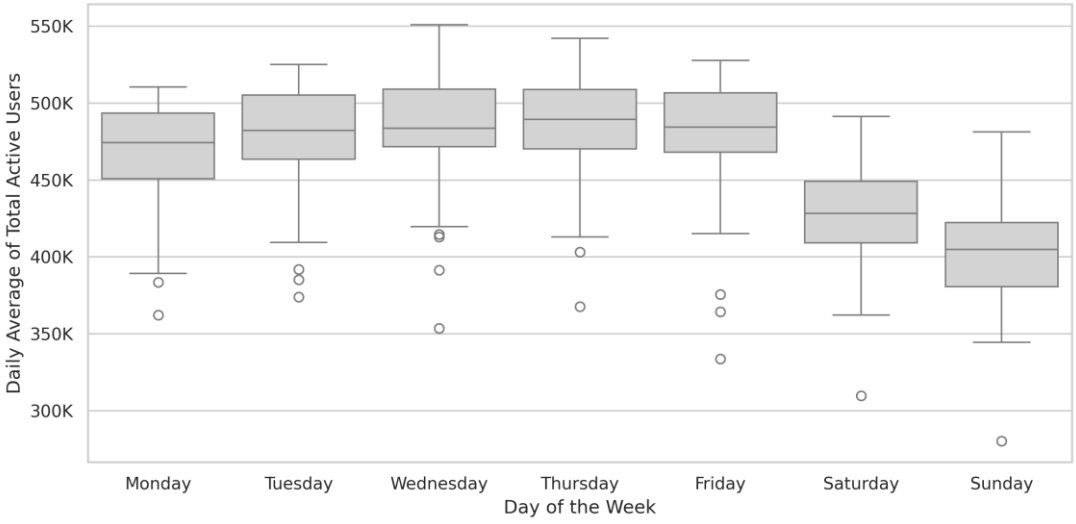

Figure 3-4 – Boxplot of Average Active Users per Hour by Day of the Week in Oeiras (2024)

Figure 3-5 provides a more detailed analysis by presenting the distribution of total active users across each hour of the day. A clear contrast is observed between daytime and nighttime activity levels. Notable peaks occur around 8:00 and 17:00, which likely correspond to typical workday start and end times. The wider interquartile ranges observed during these peak hours indicate greater variability, potentially reflecting differing user behaviors between weekdays, weekends, and holidays. Outliers are frequently observed, particularly during non-peak hours, suggesting sporadic patterns of activity outside the usual usage trends.

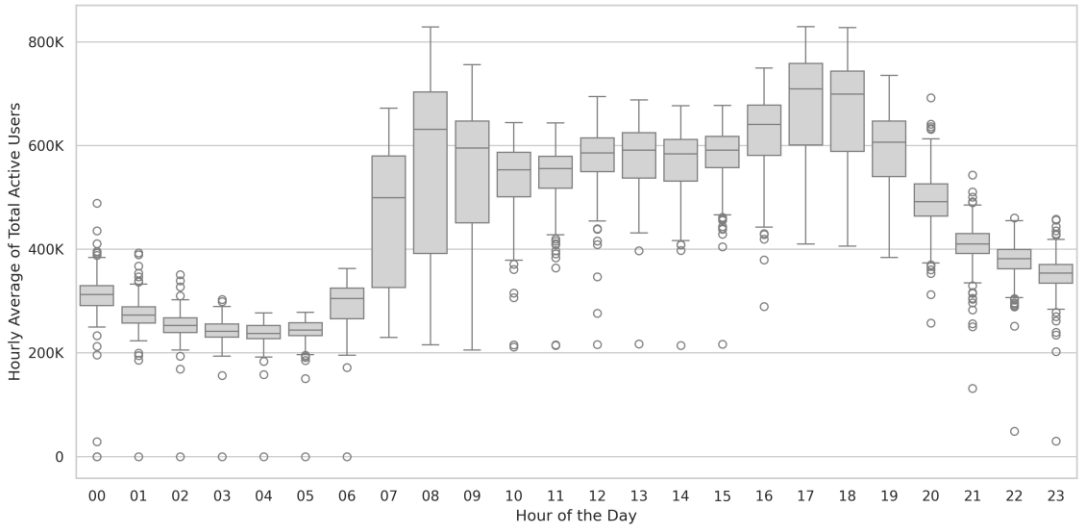

Figure 3-5 – Boxplot of Average Active Users per Hour in Oeiras (2024)

Figure 3-6 displays user activity across the statistical sections of Oeiras during daytime and nighttime periods. For the purposes of this analysis, nighttime is defined as the period between 19:00 and 06:00. Despite higher overall activity levels during the day, spatial distribution remains relatively consistent between the two periods, suggesting a possible correlation between user activity and the size or functional characteristics of each area.

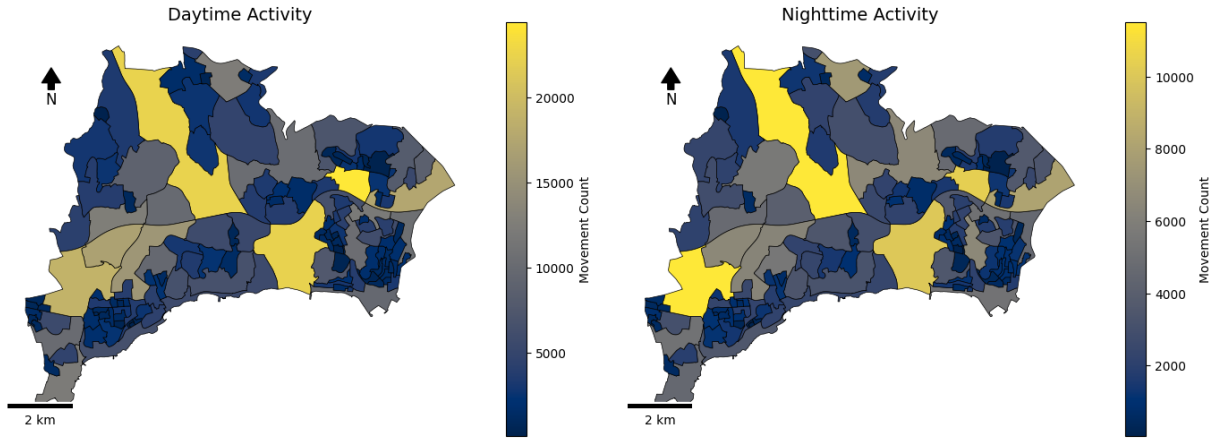

Figure 3-6 – Spatial Distribution of Active Users: Day vs. Night in Oeiras (2024)

Considering 139 statistical sections, 366 days, and 24 hourly observations per day, the dataset comprises a total of 1,220,976 records. Of these, 1,280 entries (approximately 0.1%) are

missing. Most missing values occur simultaneously across all sections on September 18th at 00:00, and on March 13th between 00:00 and 01:00. Section 111013006, which hosts operational facilities of the mobile operator, presents missing data for the entire day on February 24th, February 25th, March 2nd, March 3rd, March 9th, March 10th, and March 17th.

Given the highly dynamic nature of urban activity patterns, the presence of outliers is expected. In this analysis, outliers are defined as values that exceed 1.5 times the interquartile range (IQR), calculated individually for each statistical section. Based on this definition, 13,839 outlier cases were identified, representing approximately 1.13% of the data. These cases are mostly associated with major city events, such as summer music festivals and sports games held at the national stadium.

3.3. ORIGIN-DESTINATION MATRIX

The Origin-Destination (OD) matrix dataset, provided in CSV format, records user movements between statistical sections at two-hour intervals. The dataset includes only movements in which either the origin or the destination is located within the municipality of Oeiras. Each record contains the date of movement, the corresponding time interval, the statistical section of origin, the section of destination, and the total number of users making the trip, as detailed in Table 3-3. Movements in which the origin and destination correspond to the same section (i.e., users who remained in the same location during the observed period) are also captured.

To ensure compliance with the General Data Protection Regulation (GDPR), records involving five or fewer users are anonymized by replacing the destination section with the label "Others."

Table 3-3 – Description of Variables in the Origin-Destination Dataset

Feature	Description
datakey	Reference date for the data collection
period_range	Datetime indicating the time interval of the recorded movement
secc_origem	Code identifying the origin statistical section
secc_destino	Code identifying the destination statistical section
total	Number of active users moving from the origin to the destination

The dataset comprises 13,102,402 records, corresponding to a total of 798,025,003 user movements throughout the year 2024. Among these, 541,692 records (4.13%) represent movements in which the origin and destination sections are identical, totaling 207,745,861 movements, or approximately 26% of the overall volume.

A total of 3,500,511 records (26.7%) have anonymized destination sections labeled as "Others," as mandated by GDPR. These correspond to 85,976,890 movements, representing 10.8% of the total. An analysis of the origin sections in these anonymized records indicates that 96.5% originate from the Lisbon Metropolitan Area (LMA), and 16.2% specifically from the municipality of Oeiras. Given that the dataset includes only records where either the origin or destination is located in Oeiras, it can be inferred that all anonymized records not originating from Oeiras necessarily have Oeiras as their destination, albeit with a non-specified section.

When focusing on movements between distinct statistical sections (i.e., excluding intra-section movements), 81.92% of user movements are directed towards Oeiras. As shown in Table 3-4, 82.5% of these originate from other sections within Oeiras. The main external sources of inbound movements are Sintra (5.8%), Cascais (5.1%), Lisbon (4.4%), and Amadora (1.5%), the four municipalities that border Oeiras.

Table 3-4 – Distribution of Movement Origins with Destination in Oeiras

Origin municipality	Movements	(%)
Oeiras	398,884,666	82.5%
Sintra	27,879,215	5.8%
Cascais	24,738,059	5.1%
Lisboa	21,128,955	4.4%
Amadora	7,406,558	1.5%
Almada	1,302,731	0.3%
Loures	597,176	0.1%
Odivelas	534,500	0.1%
Vila Franca de Xira	254,975	0.1%
Others	697,430	0.2%

The remaining 18.07% of inter-sectional movements refer to cases where Oeiras is the origin. These include trips to destinations outside Oeiras or to undefined locations. Among these, 50.5% have an undefined (anonymized) destination. The main defined destinations are Sintra (20.5%), Cascais (16.0%), Lisbon (9.0%), and Amadora (3.8%), as presented in Table 3-5.

Table 3-5 – Distribution of Movement Destinations Originating in Oeiras

Destination municipality	Movements	(%)
Non-defined	53,888,311	50.5%
Sintra	21,830,471	20.5%
Cascais	17,028,083	16.0%

Lisboa	9,615,195	9.0%
Amadora	4,020,997	3.8%
Almada	267,959	0.3%
Vila Franca de Xira	8,438	0.0%
Loures	8,080	0.0%
Odivelas	7,580	0.0%
Others	13,930	0.0%

With regard to movements occurring entirely within the municipality of Oeiras, Table 3-6 presents the distribution of flows both within and between its five administrative divisions: União das Freguesias de Algés, Linda-a-Velha e Cruz Quebrada-Dafundo; Barcarena; União das Freguesias de Carnaxide e Queijas; União das Freguesias de Oeiras e São Julião da Barra, Paço de Arcos e Caxias; and Porto Salvo. In total, 84.3% of these movements take place within the same parish, indicating a high level of intra-parish mobility.

Table 3-6 – Distribution of Movement inside Oeiras

Origin	Destination	Movements	(%)
União das freguesias de Oeiras e São Julião da...	União das freguesias de Oeiras e São Julião da...	126,545,395	31.7%
União das freguesias de Algés, Linda-a-Velha e...	União das freguesias de Algés, Linda-a-Velha e...	92,982,613	23.3%
União das freguesias de Carnaxide e Queijas	União das freguesias de Carnaxide e Queijas	74,164,571	18.6%
Porto Salvo	Porto Salvo	22,166,783	5.6%
Barcarena	Barcarena	20,361,365	5.1%
União das freguesias de Carnaxide e Queijas	União das freguesias de Algés, Linda-a-Velha e...	13,020,637	3.3%
União das freguesias de Algés, Linda-a-Velha e...	União das freguesias de Carnaxide e Queijas	13,020,544	3.3%
Porto Salvo	União das freguesias de Oeiras e São Julião da...	7,268,547	1.8%
União das freguesias de Oeiras e São Julião da...	Porto Salvo	7,214,661	1.8%
Barcarena	União das freguesias de Carnaxide e Queijas	3,344,328	0.8%
União das freguesias de Carnaxide e Queijas	Barcarena	3,341,223	0.8%

União das freguesias de Algés, Linda-a-Velha e...	União das freguesias de Oeiras e São Julião da...	2,828,283	0.7%
União das freguesias de Oeiras e São Julião da...	União das freguesias de Algés, Linda-a-Velha e...	2,770,394	0.7%
Barcarena	Porto Salvo	2,407,499	0.6%
Porto Salvo	Barcarena	2,354,893	0.6%
Barcarena	União das freguesias de Oeiras e São Julião da...	1,297,136	0.3%
União das freguesias de Oeiras e São Julião da...	Barcarena	1,216,144	0.3%
União das freguesias de Oeiras e São Julião da...	União das freguesias de Carnaxide e Queijas	926,471	0.2%
União das freguesias de Carnaxide e Queijas	União das freguesias de Oeiras e São Julião da...	922,646	0.2%
Barcarena	União das freguesias de Algés, Linda-a-Velha e...	212,052	0.1%
União das freguesias de Algés, Linda-a-Velha e...	Barcarena	210,674	0.1%
União das freguesias de Algés, Linda-a-Velha e...	Porto Salvo	102,499	0.0%
Porto Salvo	União das freguesias de Algés, Linda-a-Velha e...	92,811	0.0%
Porto Salvo	União das freguesias de Carnaxide e Queijas	56,440	0.0%
União das freguesias de Carnaxide e Queijas	Porto Salvo	56,057	0.0%

Inter-parish movements are generally balanced, with similar values observed in both directions, that is, from the origin to the destination and vice versa. This symmetry is indicative of the pendular nature of daily mobility, in which individuals typically travel away from their residence and return within the same day. When considered over a sufficiently broad temporal window, such flows tend to equalize, reflecting routine commuting patterns.

Figure 3-7 illustrates the flows between parishes based on the data presented in Table 3-6. The most prominent flow, comprising approximately 13 million recorded movements, occurs between União das Freguesias de Carnaxide e Queijas and União das Freguesias de Algés, Linda-a-Velha e Cruz Quebrada-Dafundo. This is followed by flows between Porto Salvo and União das Freguesias de Oeiras e São Julião da Barra, Paço de Arcos e Caxias, with around 7.2 million movements, and between Barcarena and União das Freguesias de Carnaxide e Queijas, with approximately 3.3 million.

Except for Barcarena, these flows predominantly follow a north–south orientation, perpendicular to the direction of the Tagus River. This spatial pattern is largely influenced by the morphological characteristics of the territory, which, according to the Sustainable Mobility Plan (Município de Oeiras, 2022), is defined by the presence of watercourses and areas of steep terrain.

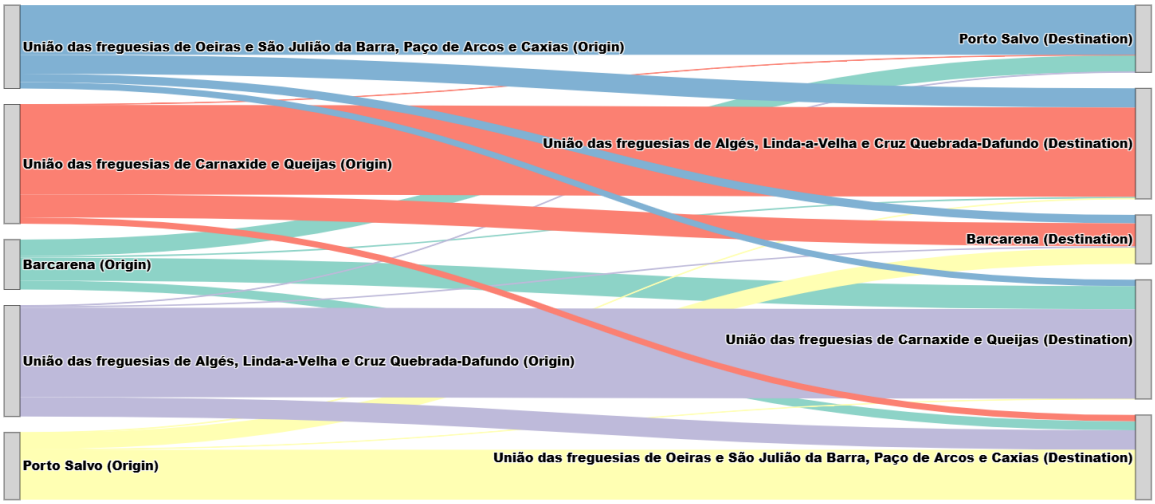

Figure 3-7 – Inter-Parish User Mobility Flows in Oeiras

4. METHODOLOGY

This chapter details the methodology used to investigate urban mobility patterns in the municipality of Oeiras through mobile phone data analysis, framing it within a Smart Mobility approach. The methodological approach aims to extract significant insights from the data to support sustainable urban planning.

The project analysis will be guided by the CRISP-DM (CRoss-Industry Standard Process for Data Mining) methodology, a widely recognized framework that structures the process into interconnected phases to ensure the robustness and clarity of the data mining project.

The data preparation stage was fundamental, ensuring that the dataset was ready for analysis. This phase included the merging of datasets, adjusting the temporal granularity of the information, normalizing features by area to mitigate biases, and handling missing values and outliers.

For the modeling phase, the study employs clustering techniques to segment city locations based on their temporal mobility patterns. Firstly, Principal Component Analysis (PCA) was applied to reduce data dimensionality and eliminate multicollinearity, optimizing the subsequent clustering process and enhancing result interpretability. Subsequently, the K-means clustering algorithm was utilized to identify latent patterns in the data, with the selection of the optimal number of clusters determined using evaluation methods such as the Elbow Method and the Silhouette Score. The visualization of the results will allow for a clear representation of the identified clusters.

4.1. CRISP-DM METHODOLOGY

The comprehensive data analysis in this study follows the CRISP-DM (CRoss-Industry Standard Process for Data Mining) methodology is a widely adopted, industry- and tool-neutral framework designed to guide data mining projects (Chapman et al., 2000). This methodology comprises six interconnected phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Chawla, 2005). The framework's provision of both structure and flexibility makes it applicable to data practitioners of varying expertise levels, from experienced analysts to those with limited technical skills (Wirth & Hipp, 2000).

CRISP-DM's effectiveness stems from its dual-layer approach, wherein its generic process model serves as a foundation for project planning, communication, and documentation, while its specialized model offers detailed steps and practical guidance for implementation (Wirth & Hipp, 2000). The methodology systematically addresses crucial aspects of data mining, including data preprocessing, exploratory data analysis, model development, and evaluation (Chawla, 2005; Jackson, 2002). Due to its adaptability and structured workflow, CRISP-DM has

become the de facto standard for technology-neutral data mining processes in both academic and industry settings (Jackson, 2002).

The application of CRISP-DM to this research is structured across multiple chapters of this thesis, reflecting the methodology's comprehensive nature. The Business Understanding phase, which focuses on defining project objectives from a research perspective and understanding domain-specific requirements, is established through the objectives outlined in the Introduction chapter and the contextual framework developed in the Literature Review chapter. This phase ensures that the technical analysis aligns with the broader research goals of investigating urban mobility patterns in Oeiras.

The Data Understanding phase, involving familiarization with the data, identification of data quality issues, and formulation of preliminary hypotheses, is comprehensively addressed in the Dataset Description chapter. This chapter provides a detailed descriptive analysis of the dataset structure, characteristics, and initial insights that inform subsequent analytical decisions. The current Methodology chapter builds upon this understanding by outlining the specific approaches for data processing and transformation into formats suitable for advanced analysis.

The Data Preparation phase encompasses the selection, cleaning, construction, integration, and formatting of data for modeling purposes. This phase is implemented through the preprocessing steps detailed in subsequent sections of this chapter, ensuring that the mobile phone data is appropriately structured for clustering analysis and pattern identification. The Modeling phase will be executed based on the analytical approaches described in the following sections, particularly focusing on the application of K-Means clustering for spatial pattern recognition.

The Evaluation phase, which assesses model performance and validates results against business objectives, will be comprehensively addressed in the Results chapter, where findings are interpreted in the context of urban mobility planning for Oeiras. The Conclusion chapter will further evaluate the research outcomes against the initial objectives and discuss the implications for sustainable urban mobility planning. The Deployment phase, while not directly applicable to this academic research, is conceptually addressed through the practical recommendations and policy implications discussed in the concluding chapters.

This structured implementation of CRISP-DM ensures a systematic and rigorous approach to the analysis of urban mobility patterns, providing a methodological foundation that supports both the technical robustness and practical applicability of the research findings.

4.2. Pre-Modeling Data Processing

The data preprocessing stage constitutes a critical phase of the CRISP-DM methodology, corresponding to the Data Preparation component. The primary objective of this phase is to

transform raw datasets into a clean, standardized format suitable for subsequent modeling techniques, particularly Principal Component Analysis (PCA) and K-Means clustering.

The preprocessing workflow involved three primary datasets: hourly records of active mobile phone users in statistical sections of Oeiras municipality, an Origin-Destination matrix of user movements recorded at 2-hour intervals, and 2021 census data including sociodemographic and geographic information. Initial data cleaning addressed format inconsistencies and standardization requirements across all datasets, ensuring compatibility for subsequent analysis phases.

To prepare the data for analysis, the available datasets were merged, combining information on active users with departure and arrival patterns for each statistical section. Since movement data was available at 2-hour intervals while user activity data was recorded hourly, temporal resolution was adjusted to ensure consistency across datasets. This alignment was achieved by filtering hourly data to retain only even-hour timestamps (0:00, 2:00, 4:00, etc.), matching the temporal granularity of the movement data. This approach preserves the temporal patterns while ensuring dataset compatibility for subsequent analysis. Using the centroids of each section, the distance of each movement was calculated and averaged to provide spatial context for mobility patterns.

A critical preprocessing decision involved normalizing all absolute-value features by their respective sections' area in square meters. This normalization was essential to account for differences in section size and prevent larger sections from biasing clustering results purely due to their geographic extent rather than their mobility characteristics. The normalization ensures that the analysis focuses on mobility intensity rather than absolute volumes, while preserving the relative patterns of movement behavior across different urban areas and maintaining the proportional relationships between different types of mobility flows.

Missing value treatment was implemented through multiple strategies depending on the data source. For Origin-Destination records with unidentified destinations, entries were filtered out, given that destination classification was essential for analysis. These missing values are primarily related to GDPR anonymization requirements for movements with fewer than 5 users, which do not significantly impact the total number of movements. During distance calculation, origins or destinations from outside the Lisbon Metropolitan Area (representing less than 0.01% of total records) were simplified by inferring distances using the maximum distance possible within the LMA boundaries. In the final merged dataset, missing movement data related to user activity was treated as zero values, which is appropriate for mobility data where the absence of records typically indicates no movement activity.

Data quality validation was performed through multiple checks, including verification of geographic coverage across all statistical sections, temporal consistency across observation periods, and logical validation of movement patterns (ensuring departure and arrival flows maintain reasonable proportions). Outlier detection and removal were implemented using the

Z-score method, which measures how many standard deviations a data point is from the mean. Values with Z-scores greater than 3 or less than -3 were classified as outliers and removed from the dataset. This threshold is widely accepted in statistical analysis as it typically identifies extreme values that fall outside 99.7% of the data distribution, assuming normality. The Z-score approach is particularly suitable for mobility data as it accounts for the natural variability in movement patterns while identifying anomalous observations that could result from data collection errors or exceptional events that do not represent typical mobility behavior.

The final processed dataset comprises features including active users per section, departure and arrival statistics, average travel distances, and region-specific movement patterns differentiated by destination and origin classifications for Lisbon, the Lisbon Metropolitan Area, and Oeiras. A detailed description of each feature in the final dataset is provided in Table 4-1. This preprocessing approach ensures a robust foundation for identifying and analyzing urban mobility patterns while maintaining analytical rigor and methodological consistency throughout the study.

Table 4-1 – Description of mobile users and movement features

Feature	Description
section_code	Section identifier based on the 2021 Portugal Census
snapshot_datetime	Timestamp of the observation period
active_users_m2	Number of active users per square meter
total_departure_m2	Total departures per square meter
avg_dist_departure	Average distance traveled in departures
lisbon_departure_m2	Departures to Lisbon per square meter
lma_departure_m2	Departures to LMA per square meter
oeiras_departure_m2	Departures to Oeiras per square meter
total_arrival_m2	Total arrivals per square meter
avg_dist_arrival	Average distance traveled in arrivals
lisbon_arrival_m2	Arrivals from Lisbon per square meter
lma_arrival_m2	Arrivals from LMA per square meter
oeiras_arrival_m2	Arrivals from Oeiras per square meter

4.3. PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) is a statistical technique used to reduce the dimensionality of large datasets while preserving as much variance as possible (Jolliffe & Cadima, 2016). It transforms the original variables into uncorrelated principal components, which are linear combinations of the original features ordered by decreasing variance. PCA

works by computing the covariance matrix of the dataset, then calculating its eigenvectors and eigenvalues (Billard & Diday, 2006).

PCA was applied to reduce feature complexity and eliminate multicollinearity, which could negatively impact the subsequent K-means clustering process. Research indicates that PCA can improve clustering performance by reducing dimensionality, which enhances computational efficiency and mitigates the curse of dimensionality (Narula, 2025; Tajunisha & Saravanan, 2010). Studies have shown that PCA improves cluster formation, interpretability, and processing time by removing noise and redundant features (Narula, 2025). This transformation is particularly valuable for mobility data, where features often exhibit high correlation due to the interconnected nature of urban movement patterns.

Before PCA implementation, all variables were standardized using z-score normalization to ensure equal contribution from each feature regardless of their original scale (Brereton, 2025). This standardization is essential when dealing with variables measured in different units, such as user counts per square meter and average distances, and prevents variables with larger scales from dominating the principal components.

To ensure an appropriate balance between dimensionality reduction and information retention, the number of Principal Components must be carefully selected. The scree plot visualization and cumulative explained variance approach were used to identify the optimal number of components. The scree plot displays eigenvalues in descending order and helps identify an "elbow point," where the marginal gain in explained variance sharply decreases. In contrast, the cumulative explained variance approach selects the smallest number of components that together explain a desired proportion of the total variance, typically 80% to 95%, and is considered more robust and interpretable in practice (Weeraratne et al., 2025). These techniques are often complemented by other methods such as the Kaiser-Guttman rule, which retains components with eigenvalues greater than 1.

Despite its advantages, PCA has limitations: it may be influenced by high-variance noise, fails to detect nonlinear structures, and later components can be difficult to interpret. However, PCA remains a powerful tool for simplifying complex datasets and increasing interpretability (Jolliffe & Cadima, 2016). The selection criteria for this study considered both the cumulative variance explained and the interpretability of the resulting components for subsequent clustering analysis, ensuring that the reduced dimensionality maintains the essential mobility patterns while facilitating effective cluster identification.

4.3 K-MEANS CLUSTERING

With the dataset appropriately pre-processed and multicollinearity addressed, the K-means clustering algorithm can be effectively employed to uncover latent patterns within the data. K-means clustering is a widely used data mining technique that partitions a set of objects into k clusters based on their similarity (Govinda Rao & Govardhan, 2014; Shukla, 2014). The

algorithm operates by partitioning observations into a predefined number of clusters, optimizing the allocation through the iterative minimization of the within-cluster sum of squared distances to each cluster centroid. This method aims to minimize the sum of squared distances between data points and their assigned cluster centroids (Govinda Rao & Govardhan, 2014), promoting the formation of compact and well-separated clusters and enabling the identification of sections and time intervals characterized by similar mobility dynamics.

K-means operates as a greedy algorithm, iteratively assigning objects to the nearest centroid and updating centroid positions (Govinda Rao & Govardhan, 2014). It is also known as nearest neighbor searching and is considered a fundamental clustering approach in data mining (Shukla, 2014). While K-means clustering is widely used, researchers have made numerous efforts to improve its performance and address its limitations (Shukla, 2014). The algorithm's simplicity and effectiveness in grouping similar objects make it a popular choice for data analysis tasks.

The subsequent step in the clustering process involves determining the optimal number of clusters to be used in the K-means algorithm. For this purpose, two widely recognized evaluation techniques are employed: the Elbow Method and the Silhouette Score. The Elbow Method involves calculating the Sum of Squared Errors (SSE) for different cluster numbers and examining the relationship between the number of clusters and the total within-cluster sum of squares (inertia), identifying a point, known as the "elbow," where the rate of decrease in SSE slows down and beyond which additional clusters provide marginal improvements in model fit (Marutho et al., 2018). This method is based on the premise that the optimal number of clusters corresponds to the point where the rate of reduction in within-cluster variance decreases significantly, forming a characteristic curvature in the graph.

In parallel, the Silhouette Score evaluates the cohesion and separation of clusters by measuring how well each point is allocated to its cluster compared to neighboring clusters. The Silhouette Score measures cluster quality based on cohesion and separation, with values ranging from -1 to 1, where higher scores indicate better clustering (Sai et al., 2017). Values close to 1 indicate that points are well-allocated to their respective clusters, values close to 0 suggest that points are on the boundary between clusters, and negative values indicate possible misallocation. Both methods have been applied successfully in various domains, including news headline categorization (Marutho et al., 2018) and crime analysis (Sagala & Gunawan, 2022). Shahapure & Nicholas (2020) demonstrated the effectiveness of using Silhouette Scores to determine and validate the optimal number of clusters in publicly available datasets.

Following the execution of the K-means algorithm, the resulting clusters will be analyzed regarding their distinctive characteristics, enabling the identification of specific urban mobility patterns. The interpretation of the groupings will be conducted through the analysis of the mean values of variables in each cluster, enabling the characterization of different traffic

regimes and mobility behaviors. This methodological approach allows not only the segmentation of data into homogeneous groups but also the discovery of valuable insights about the temporal and spatial patterns of urban mobility, contributing to a deeper understanding of transportation phenomena in the study area.

5. RESULTS

In this chapter, I will present the results obtained from the principal component and clustering analyses of the mobile aggregated geolocation data for the municipality of Oeiras. The findings reveal distinct spatial patterns across different areas of the municipality, providing valuable insights for urban planning and policy development.

Given 12 records per day over 366 days in 2024 and 139 sections in Oeiras, the final dataset comprises 610,488 rows. After removing 54,111 rows with missing values (8.9%) and 50,039 identified as outliers (8.2%), a total of 506,338 rows remain for analysis.

5.2. PRINCIPAL COMPONENT ANALYSIS

For the purposes of this analysis, selecting three Principal Components results in a cumulative explained variance of 82.4%, which represents a satisfactory balance between dimensionality reduction and information retention, as illustrated in Figure 5-1. Additionally, this choice enables effective three-dimensional visual analysis of the data.

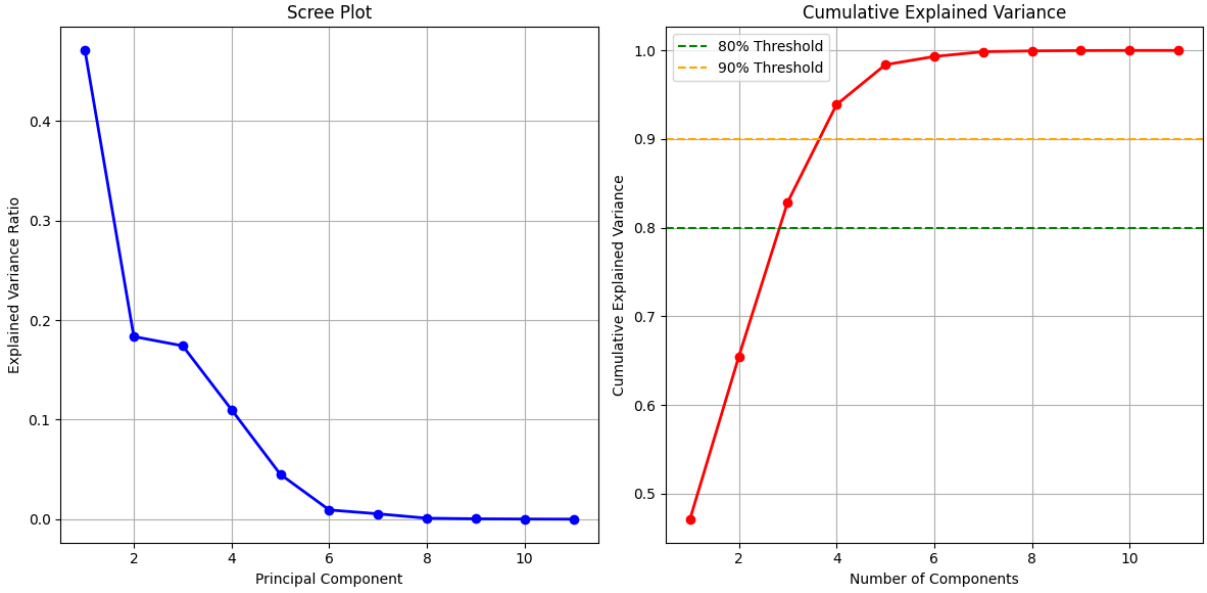

Figure 5-1 – PCA Scree Plot

After selecting the number of Principal Components and performing the linear transformation, the analysis of the component loadings provides insight into the contribution of each original variable to the resulting components. High absolute loading values indicate strong influence, with positive values denoting a direct correlation and negative values indicating an inverse relationship. The specific results of this loading analysis are presented in Figure 5-2.

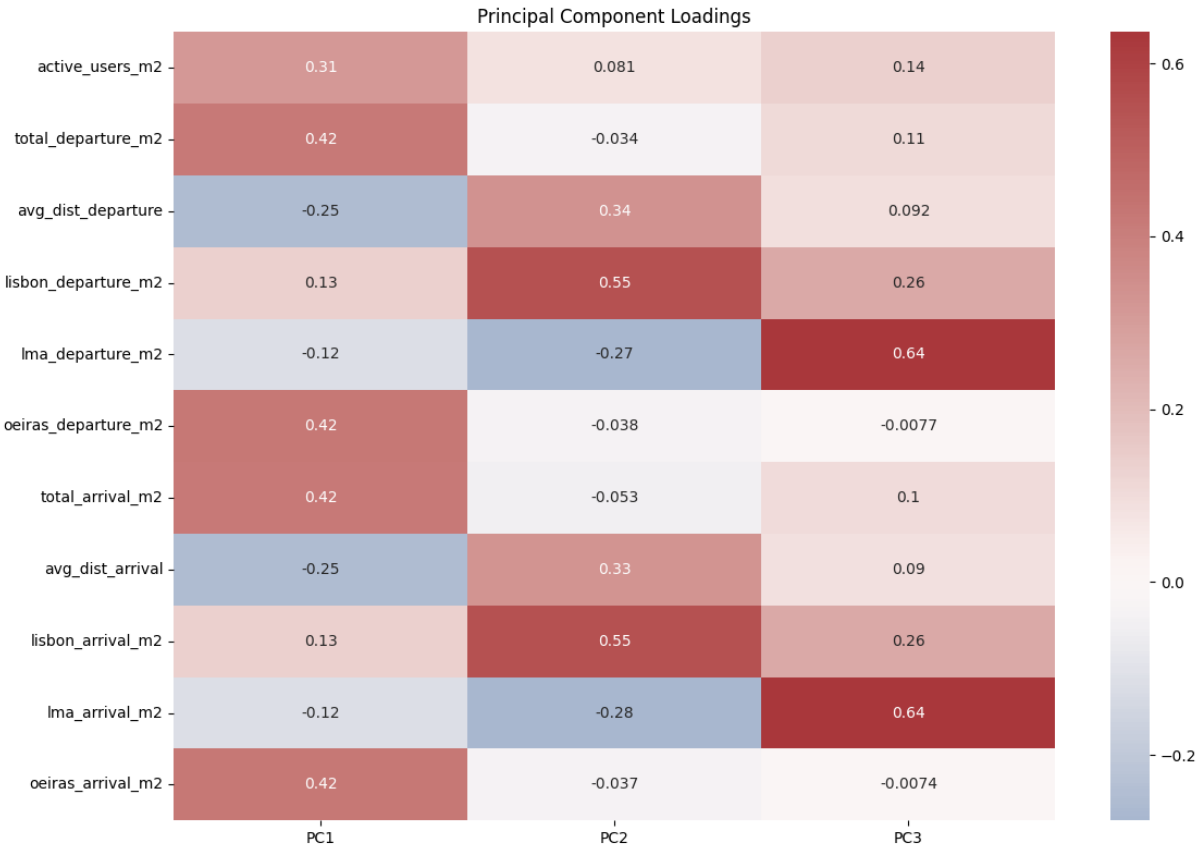

Figure 5-2 – Principal Components Loadings

Principal Component 1 (PC1) is primarily characterized by positive correlations with the variables total_departure_m2, oeiras_departure_m2, total_arrival_m2, oeiras_arrival_m2, and active_users_m2. In contrast, avg_dist_departure and avg_dist_arrival contribute negatively, albeit with comparatively smaller magnitudes. Overall, PC1 can be interpreted as representing the intensity of local user activity, with higher values associated with greater user presence and movement within sections—particularly involving Oeiras—and inversely related to longer travel distances.

Principal Component 2 (PC2) shows strong positive correlations with lisbon_departure_m2, lisbon_arrival_m2, avg_dist_departure, and avg_dist_arrival, while lma_departure_m2 and lma_arrival_m2 contribute negatively, though with smaller weights. PC2 can be interpreted as capturing long-distance movements to and from Lisbon, distinguishing them from shorter movements within the Lisbon Metropolitan Area.

Principal Component 3 (PC3) is positively correlated with lma_departure_m2 and lma_arrival_m2, and to a lesser extent with lisbon_departure_m2 and lisbon_arrival_m2. This component appears to reflect mobility patterns involving the broader Lisbon Metropolitan Area, beyond just the city of Lisbon itself.

Analyzing the results, it is evident that each principal component reflects departure and arrival behaviors similarly. This suggests that a section characterized by departures to a specific

location in the morning and corresponding arrivals from that same location in the afternoon will be similarly represented by the principal components, as the overall movement pattern remains consistent.

After reducing the dataset's dimensionality to three principal components, the data can be represented in a three-dimensional scatter plot. In Figure 5-3, each point corresponds to a section of Oeiras during a two-hour time interval in 2024.

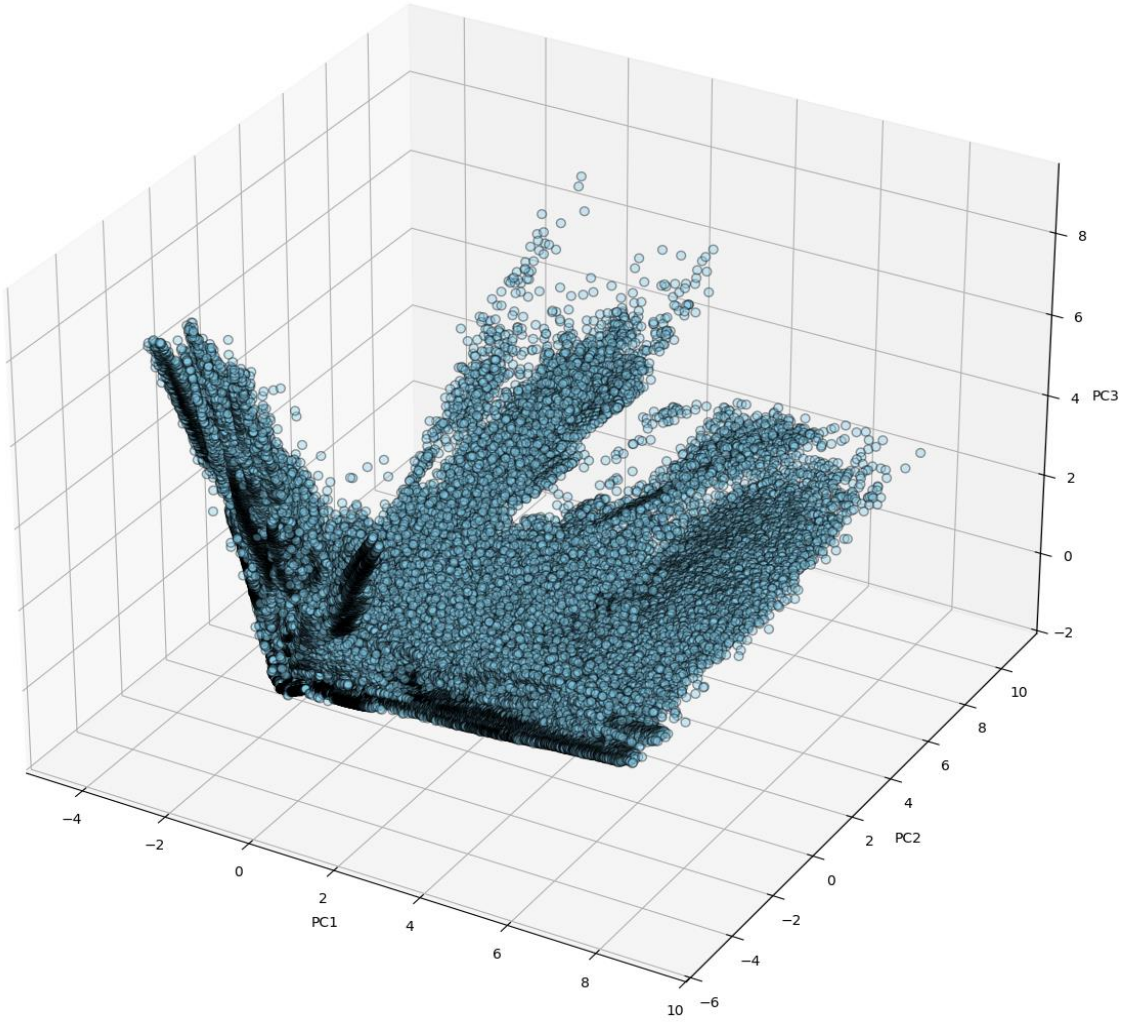

Figure 5-3 – Scatter plot of the dataset onto the first three principal components

The complex structure of the point cloud reflects diverse mobility patterns, with each nuance in its shape indicating distinct behavioral trends that will be further explored through clustering in the following step.

5.3. K-MEANS CLUSTERING

With the dataset appropriately pre-processed and multicollinearity addressed, the K-means clustering algorithm can be effectively employed to uncover latent patterns within the data.

K-means operates by partitioning observations into a predefined number of clusters, optimizing the allocation through the iterative minimization of the within-cluster sum of squared distances to each cluster centroid. This method promotes the formation of compact and well-separated clusters, enabling the identification of sections and time intervals characterized by similar mobility dynamics. The effectiveness of the clustering outcome is influenced by the initial centroid selection and the intrinsic structure of the data, underscoring the importance of prior dimensionality reduction in enhancing both the performance and interpretability of the results.

The subsequent step in the clustering process involves determining the optimal number of clusters to be used in the K-means algorithm. For this purpose, two widely recognized evaluation techniques are employed: the Elbow Method and the Silhouette Score. The Elbow Method examines the relationship between the number of clusters and the total within-cluster sum of squares (inertia), identifying a point, known as the "elbow," beyond which additional clusters provide marginal improvements in model fit. In parallel, the Silhouette Score evaluates the cohesion and separation of clusters by measuring how similar an observation is to its cluster compared to other clusters. Higher silhouette values indicate well-defined and distinct clusters. Together, these methods provide a robust foundation for selecting the optimal number of clusters, aiming to balance both model accuracy and interpretability.

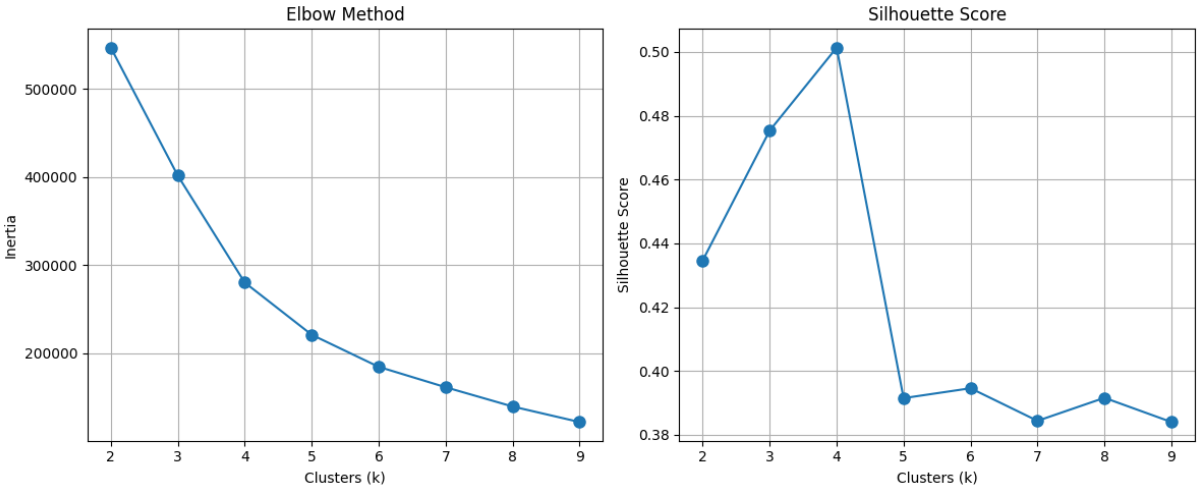

Figure 5-4 – Elbow Method and Silhouette Score

The Elbow Method yields a relatively smooth curve, in which the marginal gains in reducing within-cluster variance diminish gradually, without a clearly defined inflection point. As shown in Figure 5-4, the appropriate number of clusters could reasonably fall within the range of four to six. In contrast, the Silhouette Score exhibits a steady increase up to four clusters, after which it declines and remains at a relatively low plateau. Considering the combined evidence from both evaluation metrics, selecting four clusters represents a balanced compromise between minimizing within-cluster variance and maintaining cluster cohesion and separability.

Table 5-1 – Number of observations per cluster

Cluster	Number of Observations	Percentage
1	178,097	35,2%
2	265,094	52,4%
3	26,192	5,2%
4	36,955	7,3%
TOTAL	**506,338**	**100%**

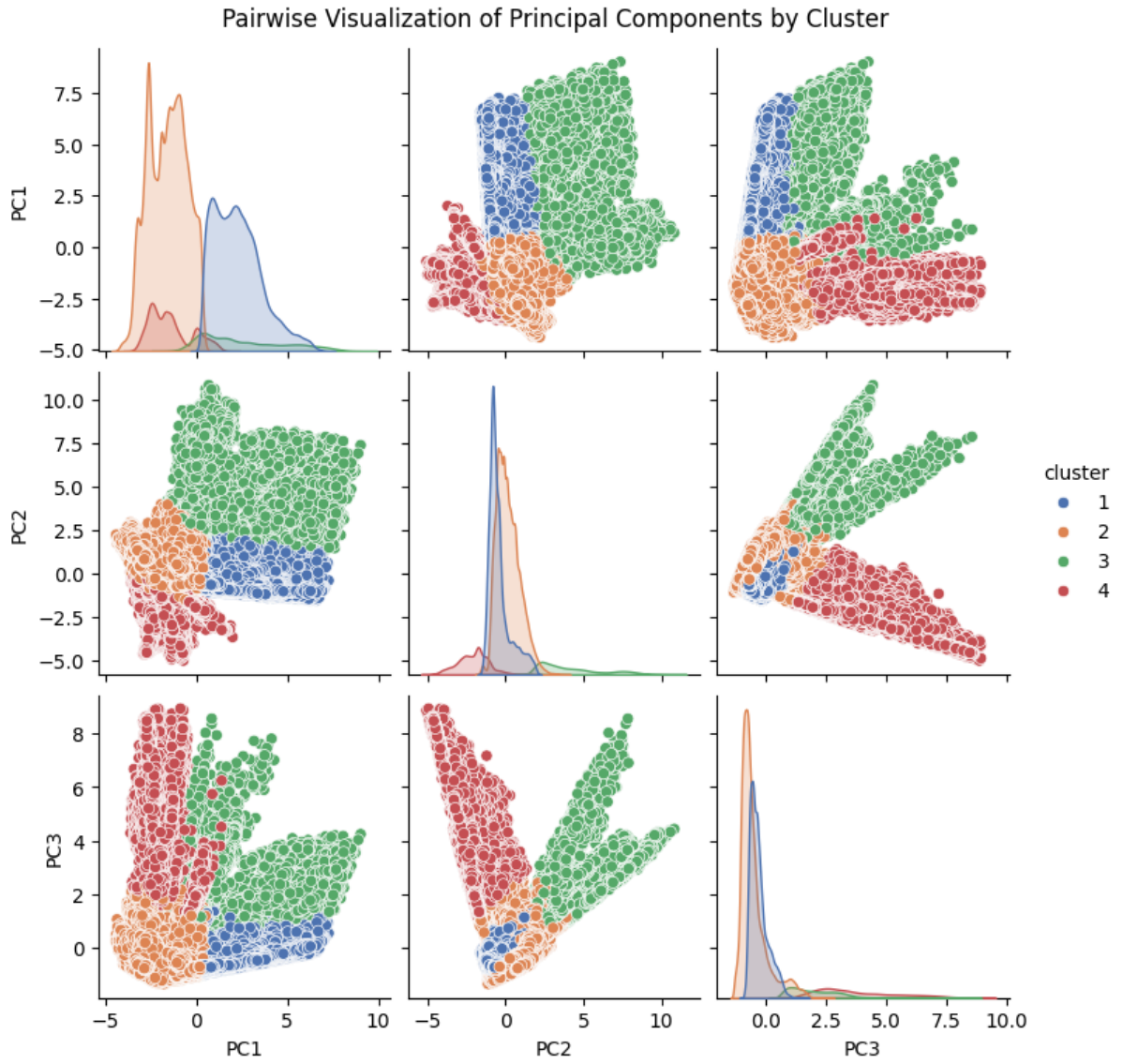

Figure 5-5 – Pairwise Visualization of Principal Components by Cluster

The application of the K-means clustering algorithm to the dataset comprising 506,338 observations yielded four distinct clusters delineated in the principal component space.

Analysis of the cluster distribution and its respective characteristics revealed noteworthy patterns in the data structure.

Table 5-2 – Cluster 1 Statistics

Cluster 1	PC1	PC2	PC3
mean	2.2833	-0.4237	-0.3070
std	1.3223	0.7017	0.3692
min	0.0446	-1.5944	-1.011
25%	1.2082	-0.8733	-0.5827
50%	2.1261	-0.6406	-0.3966
75%	3.0772	-0.2432	-0.1345
max	7.2712	2.1315	1.7175

The observations were distributed across the four clusters with varying proportions. Cluster 1 constituted the second largest segment with 178,097 observations (35.2%). This cluster exhibited a distinct pattern with substantially positive values in the first principal component (mean = 2.283), negative values in the second principal component (mean = -0.424), and slightly negative values in the t (mean = -0.307). The variability within this cluster was most pronounced in PC1 (SD = 1.322), with comparatively lower dispersion in PC2 (SD = 0.702) and PC3 (SD = 0.369).

Table 5-3 – Cluster 2 Statistics

Cluster 2	PC1	PC2	PC3
mean	-1.5894	0.1637	-0.5206
std	1.0269	0.7491	0.5728
min	-4.3882	-1.7415	-1.3300
25%	-2.4829	-0.4058	-0.8886
50%	-1.5380	0.0520	-0.6962
75%	-0.7900	0.6310	-0.3748
max	0.6327	3.9906	2.6982

Cluster 2 emerged as the predominant group, containing 265,094 observations (52.4% of the dataset). This cluster is characterized by negative values in the first principal component (mean = -1.589), marginally positive values in the second principal component (mean = 0.164), and moderately negative values in the third principal component (mean = -0.521). The standard deviations of 1.027, 0.749, and 0.573 for PC1, PC2, and PC3 respectively, indicate moderate internal homogeneity across all dimensions.

Table 5-4 – Cluster 3 Statistics

Cluster 3	PC1	PC2	PC3
mean	2.5772	4.1486	1.9954
std	2.2836	1.9783	1.0143
min	-1.5608	1.3600	0.4923
25%	0.6030	2.5251	1.1653
50%	1.9768	3.5359	1.8069
75%	4.3271	5.2800	2.6963
max	9.0360	10.8366	8.5548

Cluster 3, the smallest group comprising 26,192 observations (5.2%), presented a distinctive signature with positive values across all three principal components: PC1 (mean = 2.577), PC2 (mean = 4.149), and PC3 (mean = 1.995). This cluster exhibited the highest internal variability, particularly in PC1 (SD = 2.284) and PC2 (SD = 1.978), suggesting potentially greater heterogeneity within this group compared to the others.

Table 5-5 – Cluster 4 Statistics

Cluster 4	PC1	PC2	PC3
mean	-1.4295	-2.0728	3.8002
std	1.1450	0.9669	1.6898
min	-3.5948	-5.0462	1.1819
25%	-2.3609	-2.7267	2.4678
50%	-1.6702	-1.9871	3.3479
75%	-0.7529	-1.4289	4.9492
max	2.0049	1.3748	8.9288

Despite its relatively smaller size of 36,955 observations (7.3%), Cluster 4 demonstrated a distinctive profile. It was characterized by negative values in PC1 (mean = -1.429) and markedly negative values in PC2 (mean = -2.073), coupled with substantially positive values in PC3 (mean = 3.800). The particularly high values in PC3, ranging from 1.182 to 8.929 (SD = 1.690), constitute the defining feature of this cluster.

The four clusters identified through the analysis demonstrate clear separation in the principal component space, revealing distinct mobility patterns across the metropolitan region. This differentiation is particularly notable along the PC1 dimension, where Clusters 1 and 3 exhibited positive values (means of 2.283 and 2.577 respectively), whereas Clusters 2 and 4 displayed negative values (means of -1.589 and -1.429 respectively). This bifurcation along

PC1, which primarily represents the intensity of local user activity, constitutes a fundamental division in the dataset.

Further differentiation emerges along the PC2 dimension, which captures long-distance movements to and from Lisbon. Cluster 3 demonstrated strongly positive values (mean = 4.149), indicating substantial connectivity with Lisbon. Cluster 2 showed slightly positive values (mean = 0.164), suggesting modest Lisbon connections. In contrast, Clusters 1 and 4 exhibited negative values (means of -0.424 and -2.073 respectively), indicating limited direct Lisbon mobility patterns.

The third principal component, representing mobility patterns involving the broader Lisbon Metropolitan Area, provided additional differentiation. Cluster 4 displayed remarkably high positive values (mean = 3.800), while Cluster 3 showed moderately positive values (mean = 1.995). Conversely, Clusters 1 and 2 presented negative values (means of -0.307 and -0.521 respectively), indicating minimal engagement with broader metropolitan mobility networks.

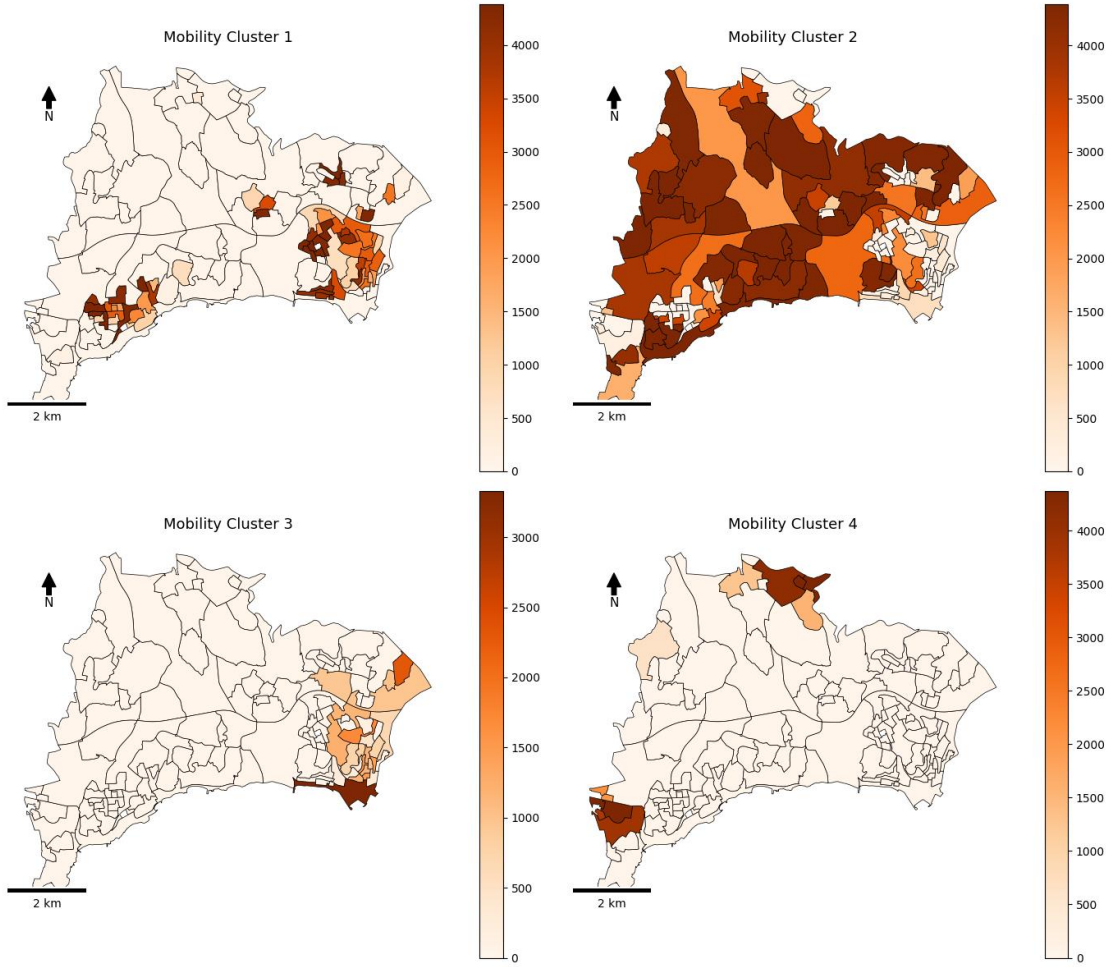

Figure 5-6 – Section Distribution by Cluster

Figure 5-6 presents the spatial distribution of clusters across statistical sections. Cluster 1 is predominantly located in areas with larger urban conglomerations. Cluster 2 represents the

most common mobility pattern and can be interpreted as a baseline behavior, occurring mainly in the central regions of Oeiras. Cluster 3 is concentrated near the border with Lisbon, while Cluster 4 appears in areas adjacent to the boundaries of neighboring municipalities.

Based on these distinctive patterns, Cluster 1 can be characterized as high local activity sections with predominant movement patterns centered around Oeiras, but limited connections to both Lisbon and the broader metropolitan area. Cluster 2 represents low-activity sections with modest Lisbon connectivity but minimal engagement with the metropolitan area. Cluster 3 sections demonstrate high local activity combined with strong Lisbon connections and moderate metropolitan area engagement. Finally, Cluster 4 exhibits low local activity but significant metropolitan area connectivity, suggesting its role as a conduit for broader regional mobility while maintaining limited internal activity.

The substantial sizes of Clusters 1 and 2 suggest they represent predominant patterns within the urban mobility landscape, while the specialized profiles of Clusters 3 and 4 indicate potentially significant subpopulations with distinct characteristics despite their smaller proportions. These findings highlight the heterogeneous nature of mobility patterns within the metropolitan region, with clear differentiation between locally oriented activity centers and those serving broader regional connectivity functions.

6. DISCUSSION OF RESULTS

6.1. URBAN MOBILITY PATTERNS IN OEIRAS

The cluster analysis based on mobile device geolocation data revealed four distinct clusters representing different urban mobility patterns in Oeiras. The results demonstrate a clear division along the PC1 dimension, which primarily reflects the intensity of local user activity. Clusters 1 and 3 exhibited positive values (means of 2.283 and 2.577, respectively), while Clusters 2 and 4 displayed negative values (means of -1.589 and -1.429, respectively).

This fundamental bifurcation in the dataset suggests a clear division between areas of high local activity (Clusters 1 and 3) and areas of low activity (Clusters 2 and 4), which aligns with observations by Benevolo et al. (2016) regarding the heterogeneity of urban mobility patterns in metropolitan contexts. These differences in local activity intensity can be interpreted as reflecting the territorial structure and land use patterns in Oeiras, as highlighted in the initial diagnosis of the municipality's Sustainable Urban Mobility and Accessibility Plan.

The differentiation observed along the PC2 dimension, which captures long-distance movements to and from Lisbon, reveals important aspects of Oeiras' integration within the broader metropolitan context. Cluster 3 demonstrated strongly positive values (mean = 4.149), indicating substantial connectivity with Lisbon, while Cluster 4 presented strongly negative values (mean = -2.073), suggesting more isolated mobility patterns in relation to the capital.

These results corroborate the observations of Gao and Zhu (2022) on how urbanization affects mobility patterns, particularly in terms of commuting distances. The identification of areas with strong connectivity to Lisbon (Cluster 3) versus areas with predominantly local mobility (Cluster 1) provides valuable insights for planning public transport infrastructure and implementing sustainable mobility strategies, as advocated by the European SUMP framework.

The third principal component, representing mobility patterns involving the broader Lisbon Metropolitan Area, provided additional differentiation between the clusters. Cluster 4 exhibited remarkably high positive values (mean = 3.800), while Clusters 1 and 2 presented negative values (means of -0.307 and -0.521, respectively).

This dimension reveals how different areas of Oeiras connect not only with Lisbon but with the entire metropolitan region. The identification of Cluster 4 as having strong metropolitan connectivity despite low local activity suggests the existence of zones that function primarily as transit points or regional connections, a pattern that had not been clearly identified in previous mobility studies in Oeiras.

6.2. COMPARISON WITH OEIRAS MOBILITY SURVEY RESULTS

When comparing our cluster analysis with the 2019 mobility survey conducted for Oeiras' Sustainable Urban Mobility Plan, we observe both significant convergences and new insights. According to the survey, residents of Oeiras made 242,580 daily trips with at least one end in the municipality. Of these, 57.3% were internal to the municipality, 26.9% connected with Lisbon, and 15.4% with other municipalities in the Lisbon Metropolitan Area (LMA), with only 0.4% extending beyond the LMA.

The cluster analysis provides complementary spatial granularity to these findings. Cluster 1, characterized by high local activity (positive PC1) and limited connectivity to both Lisbon (negative PC2) and the broader metropolitan area (negative PC3), aligns with the survey's finding that internal trips constitute the majority (57.3%) of resident mobility. The strong positive association of this cluster with Oeiras-specific mobility variables (oeiras_departure_m2, oeiras_arrival_m2) supports this interpretation.

Cluster 3, which exhibits high local activity (positive PC1) combined with strong Lisbon connectivity (highly positive PC2), corresponds to areas that facilitate the substantial volume of Lisbon-bound trips (26.9% of total trips) identified in the survey. This cluster likely represents areas where the survey found higher public transport usage (31.2% of trips), particularly "in parishes served by the Cascais line," as our geolocation data reveals these areas have stronger connectivity with Lisbon.

Similarly, Cluster 4's profile—low local activity (negative PC1) but strong metropolitan area connectivity (highly positive PC3)—aligns with the survey's finding that 15.4% of trips connect Oeiras with other LMA municipalities. This cluster likely represents areas that serve primarily as transit zones or connection points for regional movement rather than destinations themselves.

The mobility survey also revealed that Oeiras receives approximately 51,547 daily trips from non-residents, with 15,127 originating from Lisbon and 28,835 from other LMA municipalities (particularly 9,246 from Sintra). Our cluster analysis provides spatial context for these inflows, with Clusters 3 and 4 likely serving as the primary destination zones for these external trips based on their connectivity profiles.

An interesting divergence appears in the modal split data. While the survey indicates that 45.5% of residents use private vehicles (38.8% as drivers, 6.7% as passengers), 31.2% use public transport, and 20.8% walk, our geolocation-based clusters cannot directly distinguish between transportation modes. However, the inverse relationship between travel distances (avg_dist_departure, avg_dist_arrival) and local activity intensity (PC1) in our analysis suggests that areas with higher pedestrian activity may correspond to portions of Clusters 1 and 3 that exhibit high local activity.

6.3. Implications for Urban Planning and Public Policies

The results of this study have direct implications for urban planning and mobility policies in Oeiras, especially in the context of its Sustainable Urban Mobility and Accessibility Plan.

Regarding public transport optimization, the identification of areas with strong connectivity to Lisbon (Cluster 3) and the broader metropolitan area (Cluster 4) can guide the planning of more efficient public transport routes. This is particularly relevant given the survey findings that public transport represents 31.2% of resident trips but only 33.9% of trips made by non-residents working or studying in Oeiras. The average travel time of 59 minutes for external trips (compared to 33 minutes for internal trips) suggests significant potential for optimizing these connections, particularly in areas corresponding to Cluster 3. The combination of private vehicle and public transport in internal trips was found to be particularly penalized by the difficulty in making transverse connections, especially in public transport, as noted in the mobility survey.

In terms of active mobility promotion, areas of high local activity (Clusters 1 and 3) present potential for the development of active mobility infrastructure such as bicycle lanes and pedestrian areas. The mobility survey indicated that pedestrian mode already represents 20.8% of resident trips, being "particularly important for internal trips." Our cluster analysis helps identify specific zones where infrastructural improvements could further increase this percentage, targeting areas with already high local activity patterns. The current near-absence of cycling as a transportation mode (less than 1.5% according to the survey) represents an opportunity for significant growth if appropriate infrastructure is developed in these high-activity zones.

Addressing private vehicle dependency emerges as a critical challenge, with the mobility survey revealing that private vehicles account for 45.5% of resident trips and a concerning 62.8% of trips made by non-residents working or studying in Oeiras. Our cluster analysis can identify strategic intervention points, particularly in areas corresponding to Cluster 2 (low local activity but some Lisbon connectivity) and transition zones between clusters, where targeted public transport improvements might reduce car dependency. The long average distance traveled by those who work or study in Oeiras (83.7 km) further emphasizes the importance of providing viable alternatives to private vehicles.

For land use planning, the clear differentiation between areas of high and low local activity can inform decisions on zoning and urban development. The survey's finding that internal trips constitute 57.3% of resident mobility suggests that promoting mixed-use development in areas corresponding to Cluster 1 could further reduce the need for external travel, building upon existing patterns of local mobility and potentially increasing the proportion of trips that can be made on foot.

Regional connectivity strategies should consider that 15.4% of resident trips connect to LMA municipalities other than Lisbon (particularly Cascais at 6.5% and Sintra + Amadora at 5%), and 28,835 incoming trips originate from these municipalities, with Sintra alone accounting for 9,246 trips. The areas identified in Cluster 4 emerge as critical intervention points for regional connectivity initiatives that could improve the efficiency of these inter-municipal movements.

Finally, regarding emissions reduction, understanding the spatial distribution of mobility patterns can assist in developing targeted strategies for reducing greenhouse gas emissions, particularly in areas with high private vehicle usage. The travel distances identified in the mobility survey (averaging 83.7 km for non-residents) highlight the significant environmental impact of current commuting patterns. Strategic interventions in areas corresponding to Clusters 3 and 4, which demonstrate strong external connectivity patterns, could yield substantial environmental benefits by shifting these longer trips to more sustainable transportation modes.

6.4. LIMITATIONS AND CHALLENGES

While geolocation data-based analysis offers valuable insights, it is important to acknowledge several limitations and challenges inherent to this methodological approach.

The issue of sample representativeness constitutes a primary concern, as the geolocation data reflects only mobile device users from the specific operator in question. While this approach may not fully capture the demographic diversity of the population, potentially underrepresenting groups such as elderly individuals or those with lower socioeconomic status who may exhibit distinct mobility patterns or have limited access to mobile devices, the data is processed by the operator with the aim of reducing such biases. Despite these efforts, limitations in representativeness remain when compared to the systematic sampling design of the mobility survey, which explicitly sought balanced demographic coverage. An additional limitation concerns the quality of the geolocation data itself: the positional accuracy can vary significantly, and the associated error is not quantified by the operator. This introduces a degree of uncertainty that must be acknowledged when interpreting the spatial precision of movement data.

Temporal granularity presents another significant limitation. The aggregation of data at regular intervals may obscure short-duration or irregular mobility patterns that could be relevant for comprehensive mobility planning. While the mobility survey captured trip purposes and frequencies, our geolocation data provides continuous measurement, but without contextual information about the motivations behind observed movements. The survey revealed significant variations in modal choice and trip characteristics based on purpose (work, study, leisure), which cannot be directly inferred from geolocation data alone.

Cluster interpretation requires deep knowledge of the local context, and some nuances may be lost in statistical abstraction. The application of principal component analysis and clustering algorithms, while mathematically rigorous, necessitates subjective interpretation when translating statistical patterns into meaningful mobility phenomena. The process of dimension reduction inherently sacrifices some granularity in favor of pattern detection. For instance, while we can identify Cluster 3 as having strong connectivity with Lisbon, we cannot directly distinguish between the 31.2% of trips made by public transport and the 45.5% made by private vehicles, as documented in the mobility survey.

Privacy and ethical considerations remain relevant in the use of geolocation data for urban mobility analysis, even though the data was collected anonymously and aggregated. Unlike the mobility survey, which obtained explicit consent from participants, passive data collection raises questions about awareness and implicit consent. The balance between extracting valuable planning insights and respecting individual privacy rights requires continuous evaluation and adherence to evolving data protection standards.

Additionally, our analysis faces limitations in capturing modal split information that was explicitly gathered in the mobility survey. While the survey identified that 45.5% of residents use private vehicles, 31.2% use public transport, and 20.8% walk, our geolocation data cannot directly distinguish between these transportation modes. The integration of both data sources, therefore, becomes essential for comprehensive mobility planning that addresses both spatial patterns and modal choices.

Despite these limitations, the complementary nature of geolocation data analysis and traditional survey methods offers a more comprehensive understanding of mobility patterns than either approach could provide in isolation. The continuous, spatial nature of geolocation data complements the contextual richness of survey data, suggesting that future mobility planning exercises should leverage both methodologies.

6.5. Directions for Future Research

Based on the results and limitations of this study, several promising directions for future research emerge that could further enhance our understanding of urban mobility patterns in Oeiras and similar contexts.

Longitudinal analysis represents a critical next step in mobility research. Future studies could explore how mobility patterns in Oeiras evolve over time, especially in response to urban interventions or changes in transport policies. This temporal dimension would provide valuable insights into the stability of the identified clusters and their sensitivity to factors such as seasonal variations, economic fluctuations, or planned events. Tracking the evolution of mobility patterns before and after the implementation of specific SUMP measures could provide empirical evidence of their effectiveness and guide future interventions. This

approach would address a limitation of both our analysis and the mobility survey, which provide static snapshots rather than dynamic views of mobility patterns.

The integration of multiple data sources offers another fertile avenue for research. Combining geolocation data with other sources such as public transport electronic ticketing, traffic sensors, environmental monitoring stations, and socioeconomic indicators could provide a more holistic view of mobility patterns and their impacts. This multi-source approach could help bridge the gap between observed movements (as captured by geolocation data) and transportation modes (as identified in the mobility survey). Particularly valuable would be the integration of public transport usage data to understand how the 31.2% of resident trips and 33.9% of non-resident trips that use public transport align with the spatial patterns identified in our cluster analysis.

Predictive modeling represents an advanced direction for future research. Developing predictive models that can anticipate changes in mobility patterns in response to specific events, urban interventions, or transportation policy changes would provide valuable planning tools for local authorities. Machine learning approaches could leverage historical data to forecast how alterations in infrastructure, service provision, or land use might affect the distribution of movements across the identified clusters. This predictive capability would be particularly valuable for evaluating proposed SUMP interventions before their implementation, potentially increasing their effectiveness and resource efficiency.

Comparative analysis extending to other municipalities in the Lisbon Metropolitan Area would provide a broader regional context. Given that 15.4% of resident trips connect with other LMA municipalities and 28,835 daily trips enter Oeiras from other LMA locations, understanding the mobility patterns across municipal boundaries is essential for comprehensive planning. Comparing the cluster structures identified in Oeiras with those of neighboring municipalities could reveal regional mobility systems and interdependencies that transcend administrative boundaries. This regional perspective would be particularly valuable for addressing the challenges associated with the average 83.7 km traveled by non-residents working or studying in Oeiras.

Behavioral research examining the decision-making processes underlying the observed mobility patterns represents another important direction. While our analysis identifies where and when people move, understanding why they choose particular routes, times, and transportation modes requires complementary qualitative research. This behavioral dimension could help explain the factors driving the modal split observed in the mobility survey (45.5% private vehicle, 31.2% public transport, 20.8% pedestrian) and identify leverage points for shifting toward more sustainable transportation choices.

Finally, the development of interactive visualization and decision support tools based on the clustering results could enhance the practical application of our findings. Creating interfaces that allow planners and policymakers to explore the spatial and temporal dimensions of

mobility patterns interactively could facilitate evidence-based decision-making and more effective communication with stakeholders. These tools could integrate both the spatial patterns identified in our analysis and the complementary insights from the mobility survey, providing a comprehensive platform for mobility planning in Oeiras.

7. CONCLUSIONS

This research successfully employed mobile phone geolocation data to analyze urban mobility patterns in Oeiras within a smart mobility framework. The application of the CRISP-DM methodology, coupled with principal component analysis and clustering techniques, yielded significant insights into the spatial and temporal dynamics of movement within the municipality and its connections to the broader Lisbon Metropolitan Area. The analysis revealed four distinct mobility clusters, each characterized by unique patterns of local activity and external connectivity, providing a nuanced understanding of how different areas of Oeiras function within the urban mobility system.

The first research objective sought to analyze urban mobility patterns by identifying regions with similar behavior throughout the week. Through principal component analysis, we identified three key dimensions of mobility variation: PC1 representing the intensity of local user activity (particularly involving Oeiras), PC2 capturing long-distance movements to and from Lisbon, and PC3 reflecting mobility patterns involving the broader Lisbon Metropolitan Area. These dimensions successfully characterized the spatial and temporal dynamics of mobility in Oeiras, revealing how different areas serve distinct functions within the urban system.

The second objective involved implementing clustering methodologies to classify urban areas based on shared mobility characteristics. The application of clustering algorithms to the principal component space yielded four well-differentiated clusters. Cluster 1 exhibited high local activity with limited external connectivity, representing areas of intense local mobility. Cluster 2 showed low overall activity but maintained some connection to Lisbon. Cluster 3 combined high local activity with strong Lisbon connections, suggesting areas that serve as important interfaces between local and metropolitan mobility systems. Cluster 4 demonstrated low local activity but significant metropolitan area connectivity, identifying zones that primarily function as conduits for broader regional movement.

The third objective aimed to analyze the results in comparison with the findings of the previous mobility survey to identify trends and disparities. Our analysis revealed substantial alignment with the survey's findings regarding the distribution of trips (57.3% internal, 26.9% to Lisbon, 15.4% to other LMA municipalities), but provided additional spatial granularity that was not available through traditional survey methods. The identification of specific areas corresponding to different mobility patterns offers valuable insights for targeted interventions aligned with Oeiras' Sustainable Urban Mobility Plan strategies.

The integration of geolocation data analysis with traditional survey methods demonstrated the complementary nature of these approaches. While the survey provided detailed information on trip purposes, transportation modes, and user demographics, our geolocation-based analysis offered continuous spatial coverage and revealed patterns that might not be

captured through self-reported behavior. This complementarity suggests that future urban mobility planning could benefit significantly from the integration of both methodologies.

The findings of this research have several important implications for urban planning and policy in Oeiras. The identification of areas with different mobility characteristics can inform targeted interventions to promote sustainable transportation modes, optimize public transit routes, and enhance active mobility infrastructure. The clear differentiation between areas of high and low local activity can guide land use planning decisions to create more self-contained neighborhoods that reduce the need for long-distance travel. Additionally, the recognition of areas serving as important connectors to Lisbon and the broader metropolitan area can help prioritize regional transportation initiatives.

Despite the valuable insights provided by this research, several limitations must be acknowledged. The reliance on data from a single mobile operator raises questions about sample representativeness. The aggregation of data at regular intervals may obscure short-duration or irregular mobility patterns. The statistical abstraction inherent in clustering approaches necessitates careful interpretation grounded in local knowledge. Furthermore, unlike survey data, geolocation data does not directly capture transportation modes or trip purposes, limiting some aspects of analysis.

Future research could address these limitations through longitudinal studies examining how mobility patterns evolve over time, particularly in response to urban interventions or policy changes. The integration of multiple data sources, including public transport ticketing, traffic sensors, and socioeconomic indicators, could provide a more comprehensive understanding of mobility dynamics. Predictive modeling approaches could help anticipate changes in mobility patterns in response to specific events or interventions. Comparative analysis extending to other municipalities in the Lisbon Metropolitan Area would provide valuable regional context.

In conclusion, this research demonstrates the potential of geolocation data analysis for understanding urban mobility patterns and informing sustainable urban planning. By revealing the spatial structure of mobility in Oeiras and its connections to the broader metropolitan area, this study contributes to the emerging field of smart mobility and provides practical insights for the implementation of sustainable urban mobility strategies. The methodology developed and insights gained can inform similar analyses in other urban contexts, contributing to the broader goal of creating more sustainable, efficient, and livable cities.

8. BIBLIOGRAPHICAL REFERENCES

Allam, Z., & Sharifi, A. (2022). Research Structure and Trends of Smart Urban Mobility. *Smart Cities*, *5*(2), 539–561. https://doi.org/10.3390/smartcities5020029

Batra, A., & Chhabra, P. (2023). Smart cities: An empirical study of definitions. *ShodhKosh: Journal of Visual and Performing Arts*, *4*(1). https://doi.org/10.29121/shodhkosh.v4.i1.2023.357

Benevolo, C., Dameri, R. P., & D'Auria, B. (2016). Smart Mobility in Smart City: Action Taxonomy, ICT Intensity and Public Benefits. In T. Torre, A. M. Braccini, & R. Spinelli (Eds.), *Empowering Organizations* (Vol. 11, pp. 13–28). Springer International Publishing. https://doi.org/10.1007/978-3-319-23784-8_2

Billard, L., & Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining* (1st ed.). Wiley. https://doi.org/10.1002/9780470090183

Bıyık, C., Abareshi, A., Paz, A., Ruiz, R. A., Battarra, R., Rogers, C. D. F., & Lizarraga, C. (2021). Smart Mobility Adoption: A Review of the Literature. *Journal of Open Innovation: Technology, Market, and Complexity*, *7*(2), 146. https://doi.org/10.3390/joitmc7020146

Brereton, R. G. (2025). Principal Component Analysis: Standardisation. *Journal of Chemometrics*, *39*(1), e3607. https://doi.org/10.1002/cem.3607

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide* (13th ed., Vol. 9). SPSS inc.

Chawla, N. (2005). Discovering Knowledge in Data: An Introduction to Data Mining. *Briefings in Bioinformatics*, *6*(4), 411–412. https://doi.org/10.1093/bib/6.4.411

Dash, M., Koo, K. K., Krishnaswamy, S. P., Jin, Y., & Shi-Nash, A. (2016). Visualize People's Mobility—Both individually and Collectively—Using Mobile Phone Cellular Data. *2016 17th IEEE International Conference on Mobile Data Management (MDM)*, 341–344. https://doi.org/10.1109/MDM.2016.59

Demissie, M. G., Phithakkitnukoon, S., Kattan, L., & Farhan, A. (2019). Understanding Human Mobility Patterns in a Developing Country Using Mobile Phone Data. *Data Science Journal*, *18*, 1. https://doi.org/10.5334/dsj-2019-001

Dociu, M., & Dunarintu, A. (2012). *The Socio-Economic Impact of Urbanization*. *2*(1).

Fornalchyk, Y., & Hilevych, V. (2023). Characteristics of motorization's impact on the urban population. *Transport Technologies*, *2023*(2), 68–75. https://doi.org/10.23939/tt2023.02.068

Gao, Y., & Zhu, J. (2022). Characteristics, Impacts and Trends of Urban Transportation. *Encyclopedia*, *2*(2), 1168–1182. https://doi.org/10.3390/encyclopedia2020078

Goumiri, S., Yahiaoui, S., & Djahel, S. (2023). Smart Mobility in Smart Cities: Emerging challenges, recent advances and future directions. *Journal of Intelligent Transportation Systems*, 1–37. https://doi.org/10.1080/15472450.2023.2245750

Govinda Rao, S., & Govardhan, A. (2014). Assessing h- and g-Indices of Scientific Papers using k-Means Clustering. *International Journal of Computer Applications*, *100*(11), 37–41. https://doi.org/10.5120/17572-8266

Hammons, R., & Myers, J. (2019). Smart Cities. *IEEE Internet of Things Magazine*, *2*(2), 8–9.
Hasan, S., Schneider, C. M., Ukkusuri, S. V., & González, M. C. (2013). Spatiotemporal Patterns of Urban Human Mobility. *Journal of Statistical Physics*, *151*(1–2), 304–318. https://doi.org/10.1007/s10955-012-0645-0

Instituto Nacional de Estatística (Ed.). (2022). *Resultados definitivos: Portugal*.
Jackson, J. (2002). Data Mining; A Conceptual Overview. *Communications of the Association for Information Systems*, *8*. https://doi.org/10.17705/1CAIS.00819

Jolliffe, I. T., & Cadima, J. (2016a). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Jolliffe, I. T., & Cadima, J. (2016b). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Marutho, D., Hendra Handaka, S., Wijaya, E., & Muljono. (2018). The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News. *2018 International Seminar on Application for Technology of Information and Communication*, 533–538. https://doi.org/10.1109/ISEMANTIC.2018.8549751

Mitieka, D., Luke, R., Twinomurinzi, H., & Mageto, J. (2023). Smart Mobility in Urban Areas: A Bibliometric Review and Research Agenda. *Sustainability*, *15*(8), 6754. https://doi.org/10.3390/su15086754

Mohanty, S. P., Choppali, U., & Kougianos, E. (2016). Everything You Wanted to Know About Smart Cities. *IEEE Consumer Electronics Magazine*, *5*(3), 60–70.

Município de Oeiras. (2022). *Plano de Mobilidade Urbana Sustentável e Plano de Acessibilidade de Oeiras*. Gabinete de Inteligência Territorial. https://oeirasinterativa.oeiras.pt/dadosabertos/dataset/plano-de-mobilidade-urbana-sustentavel-e-plano-de-acessibilidade-de-oeiras-discussao-publica

Narula, P. (2025). Analysis of Unsupervised Clustering Algorithms and Impact of Dimensionality Reduction: A Data Driven Approach. *Machine Learning and Applications: An International Journal*, *12*(1), 169–184. https://doi.org/10.5121/mlaij.2025.12111

Noulas, A., Scellato, S., Mascolo, C., & Pontil, M. (2021). An Empirical Study of Geographic User Activity Patterns in Foursquare. *Proceedings of the International AAAI Conference on Web and Social Media*, *5*(1), 570–573. https://doi.org/10.1609/icwsm.v5i1.14175

Oginga Martins, J., & Sharifi, A. (2022). *World Cities Report 2022: Envisaging the Future of Cities.*

Porru, S., Misso, F. E., Pani, F. E., & Repetto, C. (2020). Smart mobility and public transport: Opportunities and challenges in rural and urban areas. *Journal of Traffic and Transportation Engineering (English Edition)*, *7*(1), 88–97. https://doi.org/10.1016/j.jtte.2019.10.002

Rupprecht, S., Brand, L., Böhler-Baedeker, S., & Brunner, L. M. (2019). *Guidelines for developing and implementing a Sustainable Urban Mobility Plan (2nd edition)*.

Sagala, N. T. M., & Gunawan, A. A. S. (2022). Discovering the Optimal Number of Crime Cluster Using Elbow, Silhouette, Gap Statistics, and NbClust Methods. *ComTech: Computer, Mathematics and Engineering Applications*, *13*(1), 1–10. https://doi.org/10.21512/comtech.v13i1.7270

Sai, L. N., Shreya, M. S., Subudhi, A. A., Lakshmi, B. J., & Madhuri, K. B. (2017). Optimal *K* -Means Clustering Method Using Silhouette Coefficient. *International Journal of Applied Research on Information Technology and Computing*, *8*(3), 335. https://doi.org/10.5958/0975-8089.2017.00030.6

Shahapure, K. R., & Nicholas, C. (2020). Cluster Quality Analysis Using Silhouette Score. *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, 747–748. https://doi.org/10.1109/DSAA49011.2020.00096

Shukla, S. (2014). *A Review ON K-means DATA Clustering APPROACH*.

Tajunisha, N., & Saravanan, V. (2010). An Increased Performance of Clustering High Dimensional Data Using Principal Component Analysis. *2010 First International Conference on Integrated Intelligent Computing*, 17–21. https://doi.org/10.1109/ICIIC.2010.31

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

Weeraratne, N., Hunt, L., & Kurz, J. (2025). *Optimizing PCA for Health and Care Research: A Reliable Approach to Component Selection* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2503.24248

Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a Standard Process Model for Data Mining*.

Yang, C., Zhang, Y., Ukkusuri, S. V., & Zhu, R. (2019). Mobility Pattern Identification Based on Mobile Phone Data. In S. V. Ukkusuri & C. Yang (Eds.), *Transportation Analytics in the Era of Big Data* (Vol. 4, pp. 217–232). Springer International Publishing. https://doi.org/10.1007/978-3-319-75862-6_9

Yuan, Y., Raubal, M., & Liu, Y. (2012). Correlating mobile phone usage and travel behavior – A case study of Harbin, China. *Computers, Environment and Urban Systems*, *36*(2), 118–130. https://doi.org/10.1016/j.compenvurbsys.2011.07.003

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

Universidade Nova de Lisboa