INSTITUTO
UNIVERSITÁRIO
DE LISBOA

A Data-Driven Approach to Nighttime Urban Mobility

Mariana Soares Romeu Borralho

Master's degree in Data Science

Supervisor
PhD Luís Brito Elvas, Assistant Professor
ISCTE – University Institute of Lisbon

Co-Supervised
MSc Bruno Alexandre Francisco, Invited Assistant
ISCTE – University Institute of Lisbon

October, 2025

BUSINESS SCHOOL

TECNOLOGIAS E ARQUITETURA

Department of Quantitative Methods for Management and Economics and Department of Information Science and Technology

A Data-Driven Approach to Nighttime Urban Mobility

Mariana Soares Romeu Borralho

Master's degree in Data Science

Supervisor

PhD Luís Brito Elvas, Assistant Professor

ISCTE – University Institute of Lisbon

Co-Supervised

MSc Bruno Alexandre Francisco, Invited Assistant

ISCTE – University Institute of Lisbon

October, 2025

Acknowledgement

With this thesis, I close a meaningful chapter in my pursuit of a master's in data science. The path began the day I graduated and has since grown into a steady, purposeful journey.

Over the past two years, I have had the privilege of learning from many dedicated teachers. I am especially grateful to my supervisors for their valuable guidance and continuous support, not only throughout the development of this thesis but also from the moment I joined ASIAD, where it all began.

I would also like to thank the Lisbon City Council, through the Lisbon Urban Data Laboratory, for granting access to the data that made this research possible.

Lastly, I want to express my deepest gratitude to my family and friends for their support and encouragement, which gave me the determination and resilience needed to overcome obstacles, and above all, for believing in me.

Resumo

Esta dissertação aborda o desafio de antecipar padrões complexos de mobilidade humana no contexto das cidades inteligentes, com foco na atividade noturna em Lisboa. A investigação recorre a dados de rede móvel agregados e anonimizados em larga escala, disponibilizados pela Câmara Municipal de Lisboa através do Laboratório de Dados Urbanos de Lisboa, complementados por Pontos de Interesse (POIs) georreferenciados. O objetivo central é detetar e prever variações de curto prazo na concentração populacional em áreas associadas à vida noturna da cidade. Seguindo a metodologia CRISP-DM, foi adotada uma abordagem em duas fases. Na primeira fase, aplicou-se uma análise não supervisionada para identificar e caracterizar zonas urbanas associadas à mobilidade noturna, combinando de forma complementar uma classificação estática baseada em POIs com uma clusterização comportamental suportada por PCA e K-Means. Esta abordagem permitiu selecionar 72 quadrículas representativas do fenómeno em estudo. Numa segunda fase, essas quadrículas serviram de base à formulação de um problema de regressão supervisionada, com o objetivo de prever variações de curto prazo na presença populacional. Foi conduzida uma análise comparativa de vários modelos de aprendizagem automática, incluindo algoritmos baseados em árvores e redes neuronais. O modelo MLP apresentou a melhor precisão preditiva, mantendo um desempenho robusto mesmo quando aplicado exclusivamente ao período noturno. A análise da importância das variáveis confirmou que as defasagens temporais de curto prazo são os principais determinantes da precisão do modelo. Globalmente, este trabalho contribui para uma compreensão baseada em dados da atividade urbana noturna, com implicações para o planeamento urbano e a visão da cidade inteligente de 24 horas.

Palavras-chave: Cidade Inteligente, Aprendizagem Automática, Rede Móvel, Mobilidade Noturna Urbana, Mobilidade Humana.

Códigos de classificação JEL: C45, C55.

Abstract

This thesis addresses the challenge of anticipating complex human mobility patterns in the context of smart city, with a focus on nighttime activity in Lisbon. The research leverages large-scale, aggregated, and anonymized mobile network data provided by Lisbon City Council through the Lisbon Urban Data Laboratory, complemented by geolocated Points of Interest (POIs). The main objective is to detect and predict short-term variations in population concentration in areas associated with the city's nightlife. Following the CRISP-DM methodology, a two-phase approach was adopted. In the first phase, an unsupervised analysis was applied to identify and characterize urban zones associated with nighttime mobility, combining in a complementary manner a static POI-based classification with behavioral clustering supported by PCA and K-Means. This approach enabled the selection of 72 grid cells representative of the phenomenon under study. In the second phase, these grid cells served as the basis for formulating a supervised regression problem aimed at predicting short-term variations in population presence. A comparative evaluation of several machine learning models was conducted, including tree-based algorithms and neural networks. The MLP model achieved the highest predictive accuracy, maintaining strong performance even when applied exclusively on the nighttime period. Feature importance analysis confirmed that short-term temporal lags are the main determinants of model accuracy. Overall, this work contributes to a data-driven understanding of nocturnal urban activity, with implications for urban planning and the vision of the 24-hour smart city.

Keywords: Smart City, Machine Learning, Mobile Network Data, Urban Nighttime Mobility, Human Mobility

JEL Classification Codes: C45, C55.

Index

Acknowledgement i

Resumo iii

Abstract v

List of Figures ix

List of Tables xi

List of Abbreviations and Acronyms xiii

Chapter 1. Introduction 1

1.1 Context and Motivation 1

1.2 Research Gap 2

1.3 Research Question and Objectives 2

1.4 Work Outline 3

Chapter 2. Literature Review 5

2.1 Review Methodology 5

2.1.1 Search Strategy 5

2.1.2 Inclusion and Exclusion Criteria 6

2.1.3 Results 7

2.2 Related Work 9

2.2.1 Evolution of Mobility Data and Analytical Approaches 9

2.2.2 Integration of Machine Learning and Spatiotemporal Modeling 11

2.2.3 Nighttime Urban Mobility 13

Chapter 3. Methodology 15

3.1 Business Understanding 15

3.2 Data Understanding 15

3.2.1 Mobility Data 16

3.2.2 Meteorological Data 18

3.2.3 Nightlife Indicator Identification 19

3.3 Data Preparation 19
3.3.1 Spatial Integration 19
3.3.2 POI-Based Grid Selection 19
3.3.3 Data Collection and Cleaning 20
3.3.4 Exploratory Data Analysis 20
3.3.5 Feature Engineering 26
3.4 Modeling 27
3.4.1 Behavioral Clustering of Nightlife 27
3.4.2 Data Treatment 30
3.4.3 Evaluation Metrics 31
3.4.4 Models and Parameters 31
Chapter 4. Results and Discussion 35
4.1 Overview 35
4.2 Model Performance - Global Analysis 35
4.2.1 Best Model 37
4.2.2 Permutation-Based Feature Importance Analysis 38
4.3 Nightlife - Focused Performance 40
4.4 Per-Grid Performance 41
Chapter 5. Conclusion 43
5.1 Main Findings and Conclusions 43
5.2 Limitations 45
5.3 Future Work 45
References 47
Appendix A – List of events 55

List of Figures

Figure 1 - Word cloud 6
Figure 2 - Number of studies per year 7
Figure 3 - PRISMA flow diagram 8
Figure 4 - Nightlife points of interest 20
Figure 5 - Correlation matrix 22
Figure 6 - Temporal evolution of the average number of distinct mobile terminals 22
Figure 7 - Average number of distinct mobile terminals by period 23
Figure 8 - Average hourly mobility on weekdays and weekends 24
Figure 9 - Spatial distribution on nighttime mobility intensity 25
Figure 10 – Elbow Method 28
Figure 11 - Spatial distribution of behavioral mobility clusters 30
Figure 12 - Time series of actual and predicted values using optimized MLP 37
Figure 13 - Scatter plot of actual versus predicted values with the identity line 38
Figure 14 - Top 20 feature importance 39
Figure 15 - Feature group importance 39
Figure 16 - Spatial distribution of model performance (R^2) 42

List of Tables

Table 1 - Mobile operator dataset variables 16
Table 2 - Grid dataset 18
Table 3 - Summary statistics 21
Table 4 - Top 5 nightlife grids with highest average nighttime mobility 25
Table 5 - Loadings 28
Table 6 - Clusters Results 29
Table 7 - Hyperparameter search space 33
Table 8 - Best hyperparameters 34
Table 9 - Models results 36
Table 10 - Performance metrics of the optimized MLP during the nightlife period 40
Table 11 - Summary statistics of per-grid performance metrics 41

List of Abbreviations and Acronyms

CRISP-DM	Cross-Industry Standard Process for Data Mining
LSTM	Long Short-Term Memory
LxDataLab	Lisbon Urban Data Laboratory
MAE	Mean Absolute Error
ML	Machine Learning
MLP	Multi-Layer Perceptron
MSE	Mean Squared Error
PCA	Principal Component Analysis
POI	Point of Interest
R^2	R-squared
RMSE	Root Mean Squared Error
XGBoost	Extreme Gradient Boosting

CHAPTER 1

Introduction

1.1 Context and Motivation

In recent years, the accelerated pace of urbanization and the widespread integration of digital technologies have placed the concept of the smart city at the center of urban development strategies (Kumar et al., 2020; Ferreira et al., 2023). A smart city relies on data-driven approaches to optimize infrastructure, enhance service delivery, and improve the overall quality of life for its residents (Kaluarachchi, 2022). Among the key challenges facing urban planners is the ability to understand and anticipate the complex dynamics of human mobility, particularly the flow and concentration of people across time and space (Hong et al., 2025).

Predicting where and when people concentrate in urban environments hold significant value. Accurate forecasts can inform traffic management, resource allocation, emergency response planning, and overall urban design (Ali et al., 2021). This enables public authorities to develop more sustainable, efficient, and resilient cities.

Nighttime mobility presents a distinct challenge, as movement patterns often diverge from daytime norms. Anticipating these flows is essential for ensuring public safety, managing infrastructure, and supporting local economies that thrive during evening hours (Reia, 2025). Given the dynamic and often unpredictable nature of nighttime mobility, it becomes crucial to adopt monitoring methods that offer continuous and fine-grained insights into urban flows. In this regard, data generated by mobile devices plays a strategic role (Christidis et al., 2023).

Mobile phones, as constant digital companions, generate vast amounts of location and activity data (Ghahramani et al., 2020). Mobile network operators possess anonymized and aggregated mobility records with high spatiotemporal resolution, offering a unique opportunity to model urban behavior without relying on costly and static data collection methods like manual surveys or fixed sensors (Zufiria et al., 2018).

Therefore, this work focuses on leveraging mobile network data to detect and predict short-term variations in population concentration within nightlife areas of Lisbon, enriching smart city capabilities with real-time, data-driven insight into urban dynamics during evening and nighttime hours.

1.2 Research Gap

Existing literature on urban mobility predominantly emphasizes daytime flows, commuting patterns, or movement spikes triggered by scheduled events (Bellodi et al., 2024). In contrast, research focused on regular nighttime behavior, such as spontaneous gatherings or activity around nightlife hotspots, is still limited (Alam & O'Connell, 2024; Kim 2020).

Many urban mobility studies rely solely on static spatial features like Points of Interest (POIs), without incorporating real-time behavioral indicators. This restricts the ability to anticipate and manage population surges in flexible and adaptive ways (Zhang et al., 2021; Hashemi & Zufle, 2025; Crivellari & Resch, 2022). Moreover, the influence of exogenous factors is often underexplored (Lana et al., 2018).

Given this gap in the literature, focusing on the urban area of Lisbon, the research aims to explore and model recurring nighttime mobility behavior through predictive analytics. By combining mobile network data with geolocated POIs and applying machine learning (ML) techniques, it contributes to a more data-informed understanding of nocturnal urban activity.

1.3 Research Question and Objectives

This research seeks to address the following central question: *"How can mobile network data be leveraged to analyze and predict nighttime human concentration in urban nightlife areas?"*

To answer this question, the study defines four main objectives:

O1: Analyze the existing literature to identify and assess the most relevant methodologies and algorithms used for modeling and predicting human mobility.

O2: Identify and spatially characterize nightlife zones by combining mobile network grid data with geolocated Points of Interest.

O3: Develop predictive models to forecast short-term population presence of individuals in nightlife areas.

O4: Provide insights that support data-informed urban planning and policy design.

1.4 Work Outline

The structure of this thesis is as follows: Chapter 2 presents the literature review. Chapter 3 outlines the methodology, structured according to the CRISP-DM framework. Chapter 4 presents the experimental results and discusses the model's performance and implications. Lastly, Chapter 5 concludes the study and suggests directions for future work.

CHAPTER 2

Literature Review

2.1 Review Methodology

This section outlines the methodology used to identify and select relevant literature for this research. The process followed a PRISMA literature review framework, ensuring a rigorous, transparent, and replicable review (Page et al., 2021). To ensure broad coverage and quality, the search was conducted using across two major bibliographic databases: Scopus and Web of Science.

2.1.1 Search Strategy

The search was guided by the PCC Framework (Population, Concept, and Context), which structured the identification of studies addressing urban mobility, population forecasting, and data-driven approaches to nighttime dynamics.

The search terms were grouped into three thematic categories: population-related terms (e.g., urban population, human mobility, mobile network data), conceptual terms (e.g., mobility forecasting, spatiotemporal analysis, ML and contextual terms (e.g., nightlife, urban dynamics, smart city). These keywords were combined using Boolean operators (AND/OR) to construct a comprehensive search query, applied consistently across both databases. The refined search query was as follows:

("urban population" OR "mobile phone users" OR "mobile network data" OR "people flow" OR "people density" OR "human mobility") AND ("mobility forecasting" OR "crowd prediction" OR "spatiotemporal analysis" OR "machine learning" OR "clustering" OR "geospatial modeling" OR "predictive analytics" OR "predictive modeling" OR "big data") AND ("nightlife" OR "nighttime activity" OR "evening mobility" OR "urban dynamics" OR "smart city" OR "urban planning" OR "traffic management")

To provide an overview of the thematic scope of the reviewed literature, a word cloud was generated from the titles of the studies obtained through the search query. As shown in Figure 1, the dominant terms emphasize the centrality of human mobility, urban and spatial dimensions, and the growing relevance of ML and spatiotemporal modeling in shaping current approaches to smart city research.

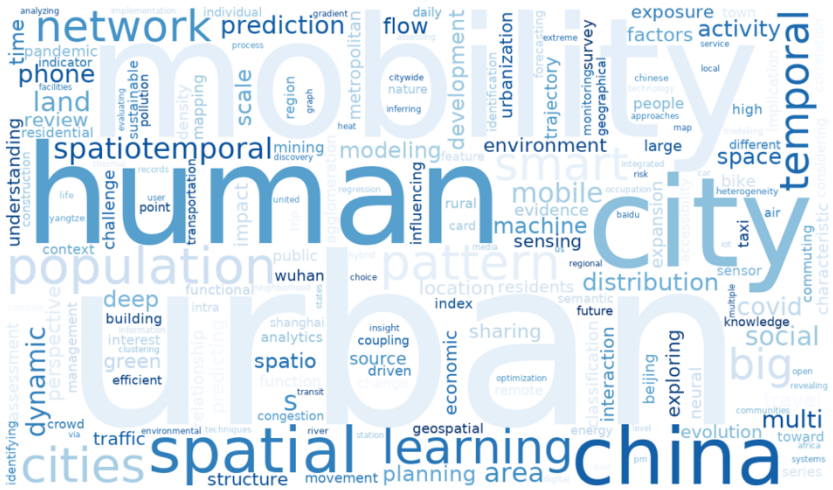


Figure 1 - Word cloud

2.1.2 Inclusion and Exclusion Criteria

To ensure the relevance, quality, and methodological consistency of the selected literature, a set of inclusion and exclusion criteria was applied. Only peer-reviewed journal articles and reviews papers published between 2019 and 2024 were considered eligible. This temporal scope was defined to capture the most recent developments and data-driven smart city research, reflecting the evolution of computational and sensing technologies over the past five years.

Studies meeting the criteria of this review were those focused on urban mobility, crowd behavior, smart city initiatives, or urban population forecasting, and demonstrating a methodological foundation in ML, spatiotemporal analysis, geospatial modeling, or mobile data analytics.

Conversely, studies were excluded if they were non-peer-reviewed publications (such as reports or conference proceedings), research outside the defined timeframe, works unrelated to urban contexts, or those lacking an empirical, analytical, or predictive component relevant to mobility analysis.

These criteria ensured that only research directly aligned with the objectives of this study was retained for review.

The growing number of publications in recent years, illustrated in Figure 2, confirms the expanding relevance of this research field. This upward trend reinforces the decision to prioritize recent literature, ensuring that the review reflects both the current state of the art and the technological advances shaping smart city research.

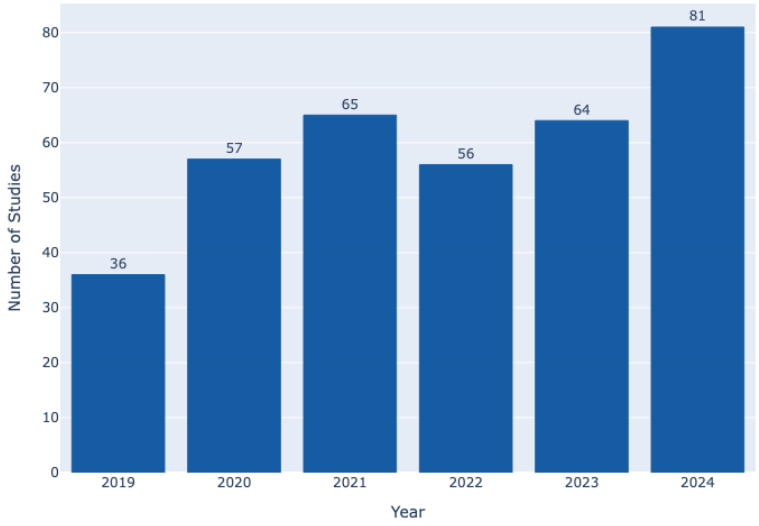


Figure 2 - Number of studies per year

2.1.3 Results

The results of the literature selection process are summarized in Figure 3, which presents the PRISMA Flow Diagram (Haddaway et al., 2022). The process followed four main stages: identification, screening, eligibility, and inclusion. Across these stages, the initial corpus of studies was progressively filtered based on predefined inclusion and exclusion criteria. The relevance of each record was assessed through a detailed review of its title, abstract, keywords, and full text. This comprehensive and structured methodology ensured transparency, reproducibility, and rigor in isolating the most relevant works aligned with the objectives of this research.

During the identification phase, a total of 449 records were retrieved from the databases described earlier (276 from Scopus and 173 from Web of Science). After removing 87 duplicate entries, 362 unique records remained for screening.

In the screening phase, titles and abstracts were reviewed leading to the exclusion of 318 records that did not meet the basic criteria or were outside the research scope.

In the following retrieval phase, 44 studies were selected for full text assessment. However, 5 of these were unavailable and consequently excluded from further analysis.

Finally, in the eligibility stage, the remaining 39 full texts were reviewed in detail to assess their methodological relevance and alignment with the research objectives. This process led to the exclusion of 20 studies that did not fully meet the eligibility requirements.

In parallel, an additional search was conducted through citation searching, allowing the identification of 4 new studies via other methods. All four were retrieved and met the eligibility criteria.

After applying all stages of the PRISMA framework, a total of 23 studies were included in the final review (19 identified through databases and 4 identified through citation searching). These publications fulfilled all the inclusion and exclusion criteria and were deemed directly relevant to the research objectives.

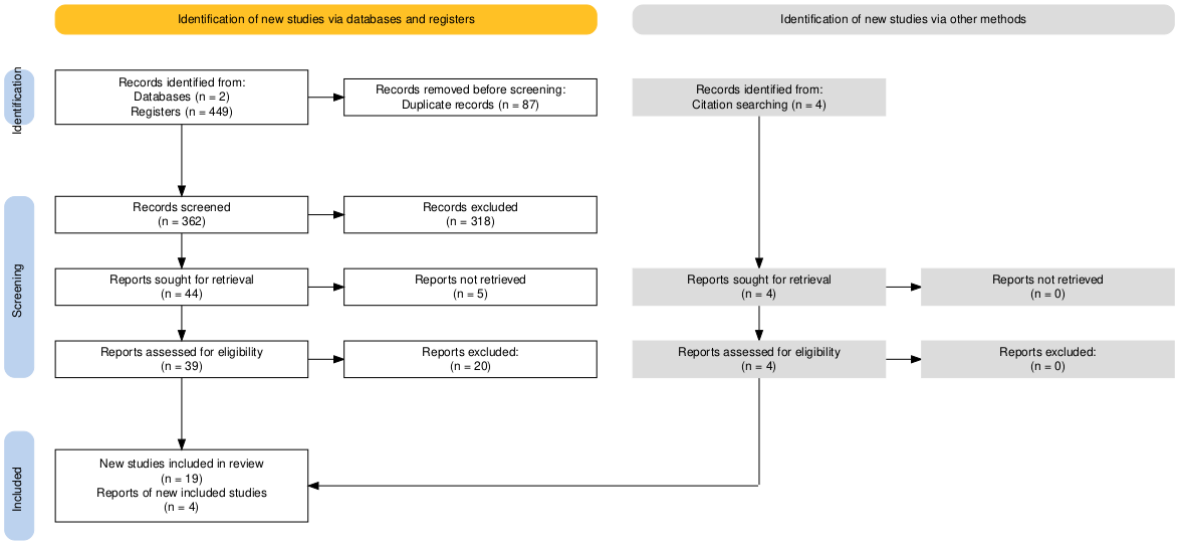


Figure 3 - PRISMA flow diagram

Based on the final set of selected publications, the following section synthesizes the main findings and methodological approaches reported in the literature. The discussion is structured around recurring themes such as mobile network data analytics, spatiotemporal modeling, ML applications, and the emerging study of nighttime urban mobility.

2.2 Related Work

The reviewed literature reveals the continuous evolution of mobility analysis in recent years, driven by the increasing availability of large-scale digital data and advances in ML. Accordingly, three major research dimensions can be identified. The first concerns the evolution of mobility data and analytical methods. The second focuses on the integration of ML and spatiotemporal modeling techniques in urban mobility prediction. Finally, the third dimension highlights the emerging attention to nighttime urban activity and its implications for smart city planning.

2.2.1 Evolution of Mobility Data and Analytical Approaches

Urban mobility research has undergone a remarkable evolution over the past decades, transitioning from static models of population distribution to dynamic, data-driven frameworks capable of capturing real-time patterns of movement. Early studies relied on census records, land-use datasets, and other static spatial features to understand urban structure and density. While these approaches provided valuable macro-level insights, they were limited in their ability to represent the temporal variability inherent to human mobility. For instance, a study integrated census data with multiple geospatial variables to improve population mapping in Guangzhou, China, using Random Forest regression. Although their model enhanced spatial accuracy, it remained constrained by the temporal rigidity of census sources (Zhao & Yang, 2020).

The advent of digital data has marked a turning point in the field. New forms of behavioral data, such as ridesharing traces, social media feeds, and location-based services, have enabled researchers to move beyond static spatial representations. In the case of ridesharing data, Roy et al. (2020) used Uber Movement data from Miami to uncover peak travel times and clustering patterns along major urban corridors, capturing short-term variations invisible in traditional data sources. For social media feeds, Santa et al. (2019) analyzed geotagged Twitter data with spatial-temporal statistical models to reveal regularities in urban activity patterns across multiple cities, highlighting the potential of social platforms for monitoring, modeling, and predicting urban dynamics. As for location-based services, Bjerre-Nielsen et al. (2020) showed that smartphone Wi-Fi and Bluetooth scans, combined with GPS, can accurately infer transportation modes (e.g., car, bus, train) with high

classification performance, demonstrating the value of passively collected device signals for urban mobility analytics.

The widespread diffusion of mobile devices has further transformed this research domain. Mobile phone data and derived products, unlike static census or survey data records offer continuous, fine-grained, and large-scale information about human presence and movement, enabling both descriptive and predictive analyses at unprecedented spatial and temporal resolutions. Wu et al. (2020) used anonymized China Unicom data at 250-meter resolution to classify urban mobility into eight subgroups, distinguishing residents, tourists, and business travelers. Wei et al. (2023) further expanded this line of research by using mobile phone user-location signal data to perform a dynamic, real-time assessment of population distribution in Xining, China. Their analysis revealed strong daily temporal regularities and spatial attenuation patterns between downtown and suburban areas.

Complementing these findings, Xu et al. (2021) focused specifically on tourist mobility, using South Korean mobile data to explore their temporal behavior through semantic labeling and time-activity diagrams to compare urban dynamics across different contexts.

In Europe, Gariazzo et al. (2019) analyzed mobile phone traffic data from over one million users in Rome, collected over a two-month period, to investigate daily mobility time profiles. Through cluster analysis, the authors identified seven distinct temporal mobility patterns corresponding to different zones and socioeconomic activities. Similarly, Bernini et al. (2019) combined mobile phone records from the Telecom Italia Big Data Challenge with land-use data in Milan to identify urban communities through cluster analysis.

The growing accessibility of heterogeneous urban datasets has led to increasing interest in data fusion approaches, which combine complementary data sources to capture a more holistic picture of urban mobility and activity.

Some recent studies have enriched mobility data analysis by integrating contextual layers such as Points of Interest (POIs). POIs help to interpret where and why people move, making mobility data more explanatory and actionable for applications such as urban planning and infrastructure management. This combination allows researchers to link movement patterns with urban functions and land use. Shi et al. (2020) combined Baidu heat maps with POI data to analyze daily travel rhythms in Wuhan, identifying systematic variations between

weekdays and weekends. Expanding this approach, Bao et al. (2023) integrated Baidu heat maps, nighttime light imagery, POI data and building volume information, a physical proxy of urban form, to estimate hourly population at a 100-meter resolution, revealing a distinct spatiotemporal pattern of "centripetal centralization at daytime, centrifugal dispersion at night".

Building further on the explicit relationship between mobility and urban form, Wang et al. (2023) combined mobile phone, POI, and land-use data to analyze congestion patterns in Xi'an using spatial clustering and Extreme Gradient Boosting (XGBoost) modeling. Their findings showed that residential and educational land-use types were the strongest contributors to congestion clusters. In a complementary study, Qi et al. (2019) adopted a more dynamic and predictive perspective, developing a multi-step framework that integrated multiple data sources, including bus smart card data and POI information, to analyze and predict regional mobility patterns in Beijing. By integrating clustering, tensor factorization, and neural network modeling, their approach captured meaningful regional mobility structures.

While these studies demonstrate meaningful progress in integrating POI information with mobility data, most research remains limited to static representations of urban space.

2.2.2 Integration of Machine Learning and Spatiotemporal Modeling

In parallel, the development of ML and spatiotemporal modeling techniques has expanded the analytical capacity of mobility studies. These methods combine spatial dependencies with temporal dynamics like daily or hourly cycles, enabling more realistic of complex urban systems. Unlike conventional regression models, ML methods can capture non-linear and high dimensional relationships inherent in urban mobility data. In this context, Bellodi et al. (2024) modeled future population density within defined census-defined subareas of Italy's Emilia-Romagna region as a regression problem. Using past density patterns within and across neighboring zones, they tested multiple ML algorithms, including tree-based ensemble methods, support vector regression, and neural networks, achieving satisfactory predictive accuracy across multiple forecasting horizons. In turn, Berlotti et al. (2024) proposed a two-level ML framework combining Time Series K-Means clustering with XGBoost to forecast

traffic flows in Catania, achieving accurate predictions even for road segments without sensors.

Beyond safety and traffic management, the integration of ML methods has increasingly been directed toward sustainable mobility objectives, aiming to optimize resource allocation, reduce emissions, and promote greener travel modes. Bertini et al. (2023), for instance, investigated cycling trips in Bologna, illustrating how temporal and spatial pattern analysis can be effectively coupled with predictive modeling for short-term demand forecasting. Their use of LSTM networks yielded higher predictive precision than other methods, demonstrating the relevance of data-driven approaches for supporting low carbon mobility planning. Extending this predictive perspective to the energy domain, Liu & Dong (2024) applied an LSTM neural network to analyze large-scale GPS data from over 93000 users in the Phoenix Metropolitan Area, successfully predicting daily urban mobility patterns with approximately 85% predictive accuracy.

These advances reflect a clear methodological shift from descriptive mapping of urban phenomena to predictive, adaptive modeling. However, the increasing availability and use of large-scale human mobility data have also raised important ethical and privacy considerations. Large scale mobility data, while invaluable for understanding urban dynamics, often contain highly sensitive information about individual behaviors and routines. As highlighted by Lin et al. (2021), even anonymized mobile phone signaling data may still reveal identifiable patterns of residential and workplace locations of users, making them vulnerable to de-anonymization. The authors emphasize that such data should therefore be regarded as personal sensitive information, requiring explicit informed consent and clear legal frameworks to ensure responsible data sharing and use in smart city contexts.

In this study, these ethical challenges were addressed by using fully aggregated and anonymized mobile network data, ensuring that all analyses focus exclusively on collective population patterns rather than individual movements. This ethical dimension is particularly relevant as smart city systems move toward real-time sensing and predictive analytics, relying on continuous streams of personal data.

2.2.3 Nighttime Urban Mobility

As smart cities evolve toward the vision of 24-hour urbanism, the concept of "nighttime economy" has gained prominence, encompassing the diverse social, cultural, and economic dynamics that characterize urban life after dark. Capturing these patterns requires analytical frameworks capable of representing their real-time spatiotemporal complexity.

Despite such progress, most mobility research focused on daytime activity, particularly commuting flows and peak hour congestion. Nighttime mobility, by contrast, represents a distinct and comparatively understudied dimension of urban life. To fill this gap, Ferreira et al. (2023) provided a large-scale analyses of nighttime mobility by applying Bidirectional Long Short-Term Memory models to mobile network data from Lisbon. Their framework predicted short-term fluctuations in population across 161 nightlife areas, achieving MAPE values below 20% and successfully identifying sudden concentration spikes.

Subsequent research has broadened this perspective by linking nighttime activity to various urban dimensions. Shi et al. (2023), for instance, developed a ML framework to identify vibrant nighttime zones in Mianyang City, integrating satellite imagery, land use and POI data to link nightlife vitality with urban regeneration processes. Lee & Lee (2024) advanced this line of inquiry by proposing a data driven approach to characterize urban nighttime vitality in Seoul. Using pedestrian traffic data and functional principal component analysis, their study extracted latent spatiotemporal components that revealed nighttime vitality patterns often overshadowed by daytime dynamics, while also highlighting seasonal and regional variations. Expanding this behavioral, Santiago-Iglesias et al. (2024) investigated the post-pandemic recovery of nightlife using Google mobility data, finding activity reductions exceeding 80% in late-night hours and highlighting the vulnerability of nightlife economies to external disruptions. Building on the socioeconomic and behavioral aspects of nighttime mobility, Kim (2020) evaluated nightlife attractiveness in Seul through an efficiency-based framework combining smart card data, open data sources, and Data Envelopment Analysis. Focusing on Millennials and Generation Z, the study quantified how demographic preferences, during of stay, and the density venues shaped the city's nocturnal hotspots.

Collectively, these studies suggest a growing recognition that urban dynamics extend well beyond conventional working hours. The city after dark constitutes a unique socio-

spatial ecosystem characterized by fluctuating population densities, distinct behavioral rhythms, and diverse economic activities. Nevertheless, systematic frameworks capable of predicting nocturnal population dynamics remain limited. Addressing this gap, the present research leverages mobile network data and ML techniques to model short-term variations in nighttime population concentration with Lisbon's nightlife areas, contributing to the broader vision of the 24-hour smart city.

CHAPTER 3

Methodology

This study follows the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, a widely used and structured framework for guiding data-driven projects (Wirth & Hipp, 2000). CRISP-DM consists of six iterative phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment, designed to ensure analytical rigor and applicability. The process is structurally iterative, allowing for a return to earlier stages whenever new insights are uncovered or unexpected challenges arise.

3.1 Business Understanding

First, understanding the real-world challenge is essential. It means identifying the specific problem the project seeks to address and ensuring that the analytical approach serves broader strategic objectives.

Urban nightlife presents unique challenges and opportunities for city planning, particularly in terms of population mobility and concentration during evening and nighttime hours. Exploring these dynamics is crucial to anticipate pressure on infrastructure, support public safety, and inform responsive urban strategies.

This project adopts a data-driven approach to analyze and predict nighttime mobility based on mobile network activity and POIs. By leveraging both telecommunications data and geolocated POIs, the study aims to model people's movements across Lisbon, especially in areas characterized by nocturnal activity.

3.2 Data Understanding

A multi-source approach was adopted to ensure a robust and multidimensional understanding of urban mobility in the Lisbon metropolitan area. By integrating diverse datasets this study constructs an analytical framework that enables spatial and temporal cross-referencing, detection of behavioral patterns, and evaluation of external influencing factors.

The following subsections detail each dataset used, including its origin, structure, and role within the overall analysis.

3.2.1 Mobility Data

The mobility dataset was provided by Vodafone, within the framework of the Lisbon Urban Data Laboratory (LxDataLab), an initiative led by the Lisbon City Council's Center for Urban Management and Intelligence. LxDataLab aims to harness the data generated in the city to produce analytical solutions that address real urban challenges and enhance the quality and responsiveness of public services for those who live, work, and visit Lisbon (Câmara Municipal de Lisboa, 2025).

By embracing advanced analytics, the project seeks to drive a new era of public service in the smart city context, unlocking deep insights from complex, large-scale datasets in areas of public interest.

In this context, Vodafone's mobility dataset offers a granular view of population dynamics across the Lisbon metropolitan area. Based on aggregated and anonymized mobile phone signals, the data spans January to December 2023, with records collected at 15-minute intervals and spatially aggregated into a grid system composed of 3743 individual 200x200 meter cells.

Table 1 provides a comprehensive overview of the variables included in the main dataset.

Table 1 - Mobile operator dataset variables

Variable	Description	Type
Grid_ID	Number of the grid	Nominal
Datetime	Date and time of the occurrence	Datetime
C1	Number of distinct terminals in the grid	Metric
C2	Number of distinct roaming terminals in the grid	Metric
C3	Number of distinct terminals that remained in the grid	Metric
C4	Number of distinct roaming terminals that remained in the grid	Metric
C5	Number of distinct terminals entering the grid	Metric
C6	Number of distinct terminals leaving the grid	Metric

Table 1 - Cont.

Variable	Description	Type
C7	Number of distinct roaming terminals entering the grid	Metric
C8	Number of distinct roaming terminals leaving the grid	Metric
C9	Number of distinct terminals with active data connection in the grid	Metric
C10	Number of distinct roaming terminals with active data connection in the grid	Metric
C11	Number of outgoing voice calls from the grid	Metric
D1	Top 10 origin countries of roaming devices	Metric
E1	Number of voice calls ending in the grid	Metric
E2	Average download speed per grid	Metric
E3	Average load speed per grid	Metric
E4	Peak download speed in the grid	Metric
E5	Peak upload speed in the grid	Metric
E6	Top 10 most used mobile applications in the grid	Metric
E7	Lowest permanence period on the grid	Metric
E8	Average permanence period on the grid	Metric
E9	Maximum permanence period on the grid	Metric
E10	Number of devices sharing mobile internet in the grid	Metric

In parallel with the mobility indicators, a complementary spatial metadata dataset was used to define and contextualize each grid cell across the metropolitan area. This dataset is outlined in Table 2.

Table 2 - Grid dataset

Variable	Description	Type
grelha_id	Grid number	Nominal
dicofre	Unique identifier of the parish	Nominal
entity_id	Unique identifier for the grid in the COI system	Nominal
entity_type	Type of entity in the COI system	Nominal
freguesia	Current parish corresponding to the grid	Nominal
freguesias	Previous parish denomination	Nominal
grelha_x	Grid position on the x-axis	Metric
grelha_y	Grid position on the y-axis	Metric
latitude	Latitude of the grid's geometric center	Metric
longitude	Longitude of the grid's geometric center	Metric
nome	Designation of the area	Nominal
objectid	Internal object identifier	Nominal
position	Spatial index within the grid dataset	Nominal
wkt	Geometric representation well-known text format	Nominal

3.2.2 Meteorological Data

Historical weather data was incorporated to represented environmental variables potentially associated with patterns of human movement.

The dataset was retrieved from the Open-Meteo Archive API (Open-Meteo, 2025), covering the full period from January 1 to December 31, 2023, for Lisbon's geographic coordinates. An API query was implemented to retrieve hourly values for the following meteorological indicators: Temperature (ºC), Precipitation (mm) and Wind Speed (km/h).

These hourly weather indicators enable to explore how different atmospheric conditions impact the volume and distribution of urban mobility (Böcker et al., 2013).

3.2.3 Nightlife Indicator Identification

To identify urban zones potentially associated with nightlife activity, spatial infrastructure indicators were analyzed. POIs tagged as bars, nightclubs, and pubs were extracted from OpenStreetMap (OpenStreetMap contributors, 2025).

However, relying solely on static infrastructure may not capture the full picture of nocturnal activity. Therefore, an unsupervised clustering approach based on actual mobility behavior was introduced to complement this static layer.

This dual perspective approach enables a more nuanced identification of nightlife zones, accounting for both formalized entertainment infrastructure and dynamic patterns of human presence during the evening period.

3.3 Data Preparation

To ensure a robust foundation for the analytical phase, the raw datasets underwent a comprehensive preprocessing workflow.

3.3.1 Spatial Integration

The preparation steps began with loading the Vodafone grid dataset, which contains spatial references encoded in well-known text format, a standard defined by the Open Geospatial Consortium for representing vector geometries in textual form (Open Geospatial Consortium, 2011). These textual geometries were cleaned to enable accurate parsing and then transformed into a spatial object using GeoPandas, resulting in a GeoDataFrame with a uniform coordinate reference system (Wasserman et al., 2020).

3.3.2 POI-Based Grid Selection

Identified nightlife POIs were reprojected to match the grid's coordinate system, allowing for spatial joins between POIs and grid geometries. Each grid cell was then enriched with both the total count of nightlife venues and a boolean flag with a minimum threshold of one venue.

A total of 227 grid cells were identified as nightlife-active, defined by the presence of one or more nightlife POIs within their boundaries. These cells are predominantly concentrated in historically areas, such as Cais do Sodré, Bairro Alto, and Santos, all of which are widely recognized as central hubs of Lisbon's evening culture. For example, the grid 403 (Cais do Sodré) contains 22 venues, showing its significance within the city's nocturnal ecosystem. The spatial distribution of nightlife hotspots is visually represented in Figure 4.

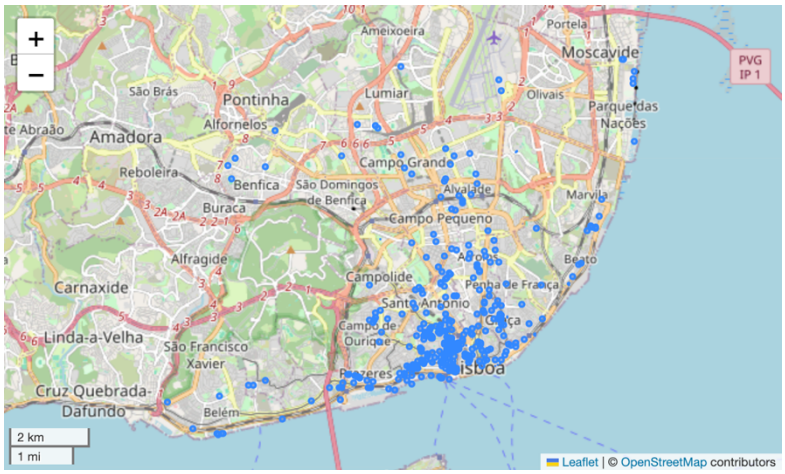


Figure 4 - Nightlife points of interest

3.3.3 Data Collection and Cleaning

The mobility records for the year 2023 were filtered to retain only entries within the selected grid cells, to isolate mobility data specific to nightlife-active locations. Monthly datasets were sequentially loaded and concatenated, resulting in a unified dataset comprising 7,868,946 records and 24 variables.

Duplicate rows were removed to ensure data integrity, and variables with missing values or irrelevant to the analytical scope were eliminated during the cleaning process.

3.3.4 Exploratory Data Analysis

This section presents an exploratory analysis of the mobile network dataset to identify preliminary patterns and temporal-spatial regularities of nighttime mobility in Lisbon's nightlife areas, using the 227 grids identified in Section 3.3.2. The analysis aims to provide descriptive insights into population dynamics before applying predictive modeling techniques.

A subset of nine numeric mobility indicators was selected for in-depth analysis to better understand their individual characteristics and interrelationships. The descriptive statistics of each variable are presented in the Table 3. Most indicators exhibit right-skewed distributions, with means consistently higher than medians and extreme maximum values. This indicates the presence of long-tailed distributions.

Table 3 - Summary statistics

Variable	Mean	Std	Min	25%	50%	75%	Max
C1	585.563	576.570	0.0	207.65	421.61	786.10	115636.50
C2	189.404	301.124	0.0	34.62	95.15	222.64	115636.50
C3	392.392	403.749	0.0	129.91	273.63	534.71	115636.50
C4	124.674	210.263	0.0	21.10	60.08	147.89	115636.50
C5	321.499	342.765	0.0	106.21	226.23	417.91	15282.92
C6	295.893	304.571	0.0	102.17	209.85	382.88	14682.68
E7	0.164	3.236	0.0	0.00	0.00	0.00	300.00
E8	24.011	35.639	0.0	11.00	13.00	15.00	300.00
E9	180.513	120.538	0.0	57.00	203.00	300.00	300.00

These indicators were visualized using a correlation matrix presented as a heatmap (Figure 5), which allowed to identify potential multicollinearity and uncover underlying dependencies between metrics. The heatmap reveals a strong internal structure among C indicators, with multiple variables such as C1, C3, and C6 showing near-linear correlations, suggesting they capture overlapping aspects of mobility behavior. In contrast, permanence indicators (E7-E9) exhibit weaker correlations with the core C metrics, reflecting distinct behavioral dimensions.

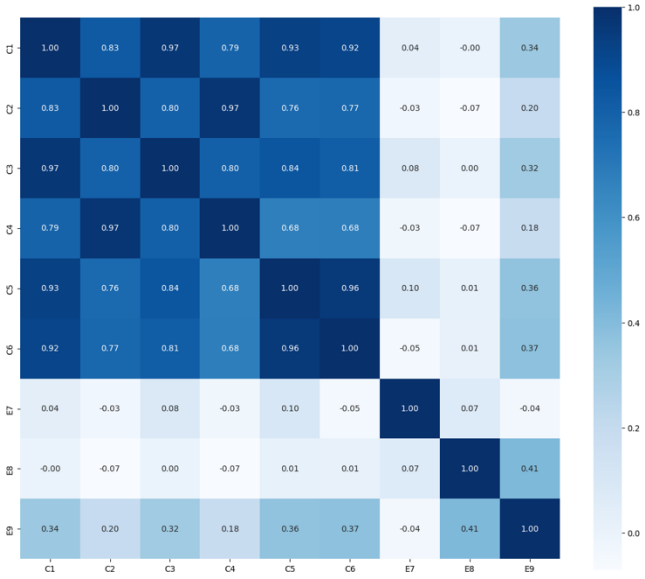


Figure 5 - Correlation matrix

To further explore temporal dynamics, the analysis examined variations in mobility across different time scales, from annual to daily and hourly patterns.

Figure 6 illustrates the temporal evolution of the average number of distinct mobile terminals in the grid throughout 2023. A clear seasonal trend emerges, with periodic fluctuations and a particularly sharp spike is observed on September 3rd at 22:30. Upon closer inspection, this peak predominantly stems from grids located in the parishes of Misericórdia and Santa Maria Maior, areas characterized by nightlife infrastructure.

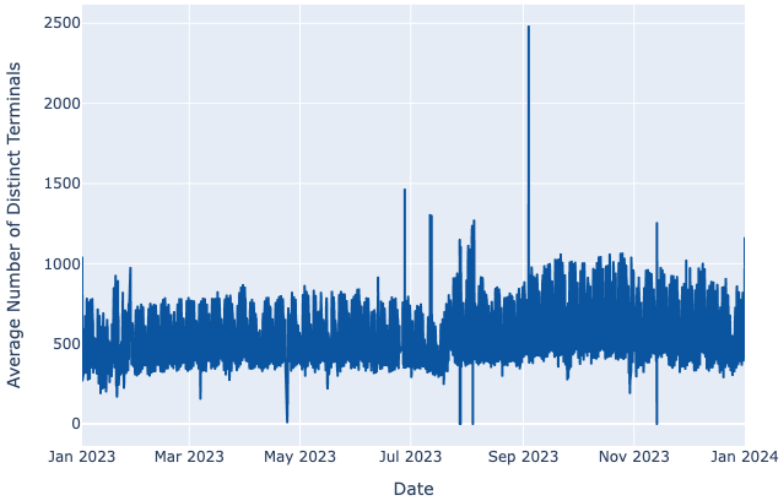


Figure 6 - Temporal evolution of the average number of distinct mobile terminals

After analyzing annual variability, the focus shifted to daily and hourly mobility rhythms. Figure 7 compares average population density between daytime and nighttime periods. Daytime activity (06:00-17:59) registers a higher average number of distinct mobile terminals in the grid (), meaning that, on average, each grid contains around 636 unique devices. In contrast, nighttime periods (18:00-5:59) record a lower average of , representing a 16% decrease after dark. Although overall population density peaks during the day, nightlife zones retain substantial activity throughout the night, underscoring their continuous role as social and economic attractors within the urban fabric.

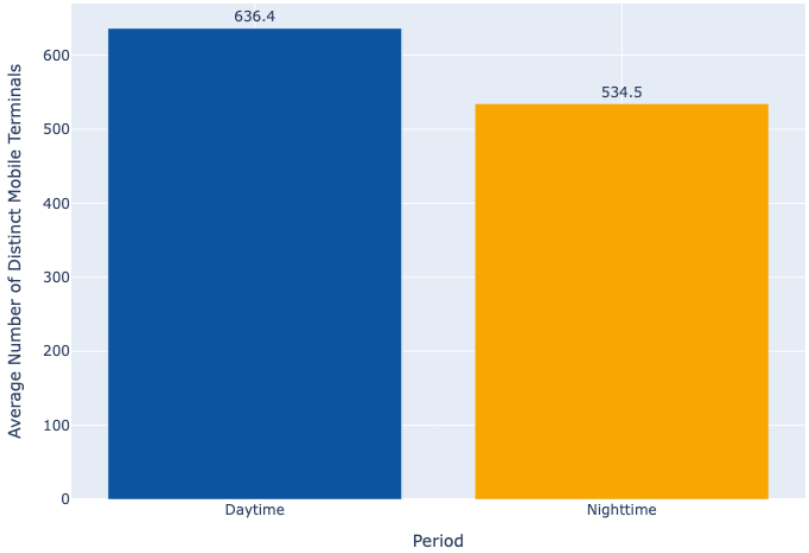


Figure 7 - Average number of distinct mobile terminals by period

The hourly evolution of population density in nightlife grids, contrasting weekdays and weekends, is shown in Figure 8. A distinct diurnal rhythm emerges, on weekdays, mobility peaks around 13:00, reflecting daytime commercial and work-related activity, followed by a gradual evening decline. In contrast, weekends display a delayed and sustained increase in nighttime activity. While on weekdays population density tends to decline around 17:00-18:00, on weekends it begins to rise, reaching its first evening peak near 21:00. Activity then temporarily declines around 23:00, before surging again to its highest level around midnight. Afterward, mobility gradually decreases until approximately 06:00, yet remains consistently higher than weekday levels, as shown by the red curve staying above the blue one throughout the night and early morning hours.

This sustained nocturnal activity reflects the extended social and leisure patterns that characterize Lisbon's nightlife areas during weekends.

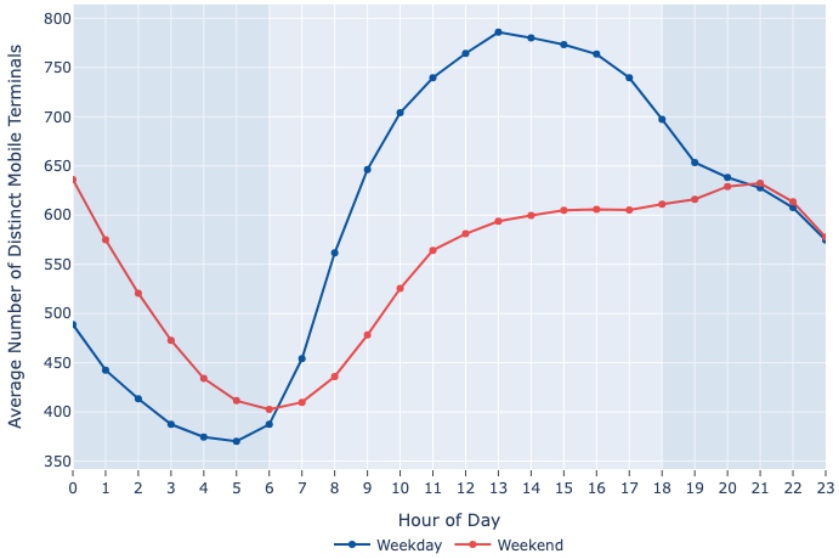


Figure 8 - Average hourly mobility on weekdays and weekends

Building upon the POI-based grid selection described in Section 3.3.2, the grids previously identified as nightlife active were analyzed to assess their actual mobility performance during nighttime hours. Figure 9 visualizes this behavioral dimension by mapping the average number of distinct mobile terminals across all nightlife grids.

The resulting spatial pattern reveals a partial overlap between infrastructure-based nightlife zones and observed mobility intensity. Several grids with few POIs exhibit high mobility, while others with dense nightlife infrastructure show moderate levels. For instance, Grid 625 in Bairro Alto with only three nightlife POIs, records an average nighttime mobility of approximately 2500 distinct terminals, whereas Grid 403, containing 22 venues, shows lower intensity (). This divergence demonstrates the limitation of relying solely on POI density and motivates the use of behavioral clustering as complementary analytical perspective in the next section.

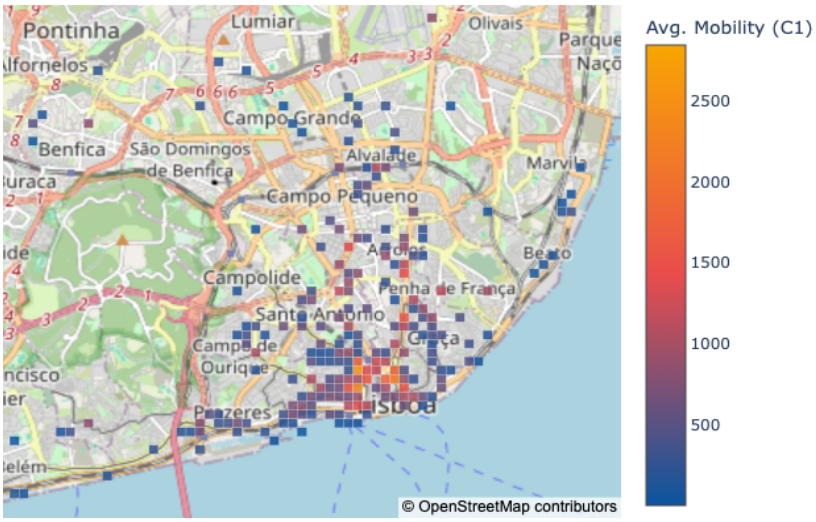


Figure 9 - Spatial distribution on nighttime mobility intensity

To provide a more detailed understanding of the spatial distribution, the analysis was extended to identify the specific grids exhibiting the highest levels of nighttime mobility. This quantitative perspective highlights the urban areas where behavioral intensity is most concentrated. Hence, Table 4 complements this spatial representation by detailing the top 5 nightlife grids with the highest average nighttime mobility. The results confirm the dominance of the historic center, where Bairro Alto (Grid 625) and Camões (Grid 512) in the parish of Misericórdia, along with Baixa (Grids 572 and 571) in Santa Maria Maior, record the strongest mobility intensities.

Table 4 - Top 5 nightlife grids with highest average nighttime mobility

Grid	Parish	Name	Average Mobility
625	Misericórdia	Bairro Alto	2538.7
512	Misericórdia	Camões	2337.3
572	Santa Maria Maior	Baixa	2169.9
571	Santa Maria Maior	Baixa	1774.1
683	Misericórdia	Bairro Alto	1755.5

To further understand how these zones evolve across time, a comparative analysis of daytime and nighttime mobility variation was conducted. This assessment quantifies how specific areas intensify or decline in population density after dark. Overall, 82% of the nightlife grids exhibited a reduction in mobility during nighttime hours, reflecting the general attenuation of daytime commuting activity. In contrast, a smaller subset (approximately 18%) displayed a marked nighttime intensification, most notably in Bairro Alto (Grid 625) and Camões (Grid 512), where mobility increased by 39% and 26%, respectively, compared daytime levels. Conversely, commercially oriented zones such as Campo das Cebolas (Grid 463) experienced substantial declines in activity, with reductions exceeding 30%, underscoring the contrasting spatial-temporal dynamics between entertainment-oriented nightlife areas and functionally daytime urban areas.

3.3.5 Feature Engineering

The feature engineering process enhances the dataset with multifaceted contextual information, supporting more robust and context-aware analytical models.

To achieve this, different types of external and derived features were incorporated. First, hourly weather data was integrated using an as-of temporal merge, to enrich the mobility dataset with temporally aligned atmospheric context. This technique matches each mobility record to the most recent preceding weather observation, ensuring temporal coherence and minimizing alignment errors.

Subsequently, to derive deeper insights from the available datasets, a suite of dummy variables was then engineered to capture key factors that could influence patterns of urban mobility.

- Nightlife period (18:00-05:59): Signaling time window typically associated with evening and nighttime social activity (Son et al., 2023).
- Day of week indicator: Differentiating behavioral patterns across working days and rest periods.
- Public holidays and their eves: Marking official holidays and their preceding days.
- Major urban events: Flagging dates of events.

- Meteorological conditions: Including weather-related variables such as rain, temperature extremes (hot/cold), and wind.

3.4 Modeling

This chapter presents the modeling approach developed to forecast short-term behavioral mobility in urban nightlife areas. The aim is to predict people's movement for the next 15-minute interval, using a combination of temporal, contextual, and environmental features.

3.4.1 Behavioral Clustering of Nightlife

Based on the dual-perspective framework introduced earlier, this subsection focuses on the dynamic component of nightlife zone identification. To complement the static infrastructure-based approach, a behavioral clustering methodology was applied to mobility indicators, covering all grid cells. The analysis specifically targeted the nighttime interval to capture activity patterns characteristic of nocturnal urban behavior.

For this purpose, monthly mobility datasets from January to December 2023 were aggregated, resulting in a unified database comprising over 63,9 million records across 3743 grid cells. Only relevant features were retained to ensure consistency. Each record was time-stamped and filtered to include only nighttime hours (between 18:00 and 5:59), isolating the temporal window most representative of nocturnal dynamics. Duplicate observations were removed prior to modeling, yielding a cleaned dataset optimized for unsupervised learning.

The clustering procedure employed the K-means algorithm, implemented through the scikit-learn library (Pedregosa et al., 2011).

Following the correlation analysis presented earlier, which revealed strong interdependencies, particularly within the C-series variables, a dimensionality reduction step was introduced to mitigate redundancy and enhance interpretability. Flows and permanence metrics were computed per grid and standardized to enable cross-feature comparison. These standardized indicators were then subjected to Principal Component Analysis (PCA), which reduced the dataset's dimensionality while preserving its main behavioral structure. The first two principal components, capturing the majority of variance, were retained for both clustering and visualization, providing a compact representation of nighttime mobility patterns across Lisbon's grids.

Table 5 presents the PCA loadings, revealing two dominant behavioral axes:

- PC1 captures overall intensity of nighttime movement.
- PC2 reveals contrasts in permanence behavior.

Table 5 - Loadings

Variable	PC1	PC2
C1	0.389042	-0.065906
C2	0.346887	-0.310377
C3	0.377472	-0.074085
C4	0.346517	-0.320478
C5	0.392433	-0.063402
C6	0.387927	-0.042541
E7	-0.224766	0.387091
E8	0.124298	0.638478
E9	0.308700	0.477289

To determine the optimal number of clusters, the Elbow Method was employed, indicating k=3 as the most suitable value, since the within-cluster cost decreases sharply up to this point and then begins to level off. The corresponding plot is presented in Figure 10.

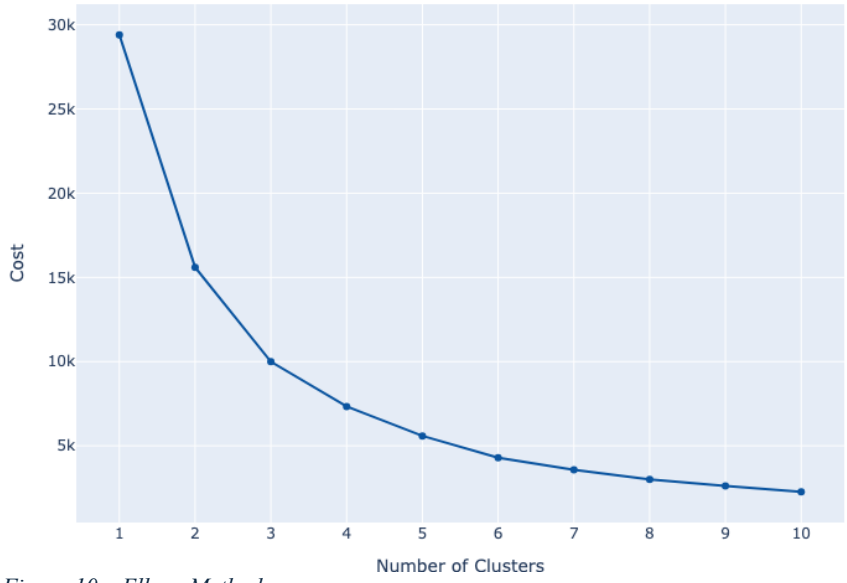


Figure 10 – Elbow Method

The K-Means algorithm was then applied to PCA transformed data, segmenting grid cells into three behaviorally distinct clusters.

Cluster 2 emerged with consistently higher values across indicators, indicating heightened nighttime activity characterized by frequent and prolonged human presence. As shown in Table 6, grids within this cluster displayed elevated values in C1, C3, C5 and C6, representing, respectively, the number of distinct mobile terminals present, the number of terminals that remain within the grid, and the counts of distinct entries and exits. This combination reflects a strong and continuous circulation of individuals within these areas during nighttime hours. Additionally, permanence related indicators such as E8 (average permanence duration) and E9 (maximum permanence duration) also recorded higher values, indicating that users tend to stay longer once present in these locations. Together, these dynamics reveal zones characterized by both high mobility turnover and prolonged occupancy. A total of 199 grid cells were assigned to cluster 2, offering a more nuanced behavioral perspective that goes beyond POI-based classifications.

Table 6 - Clusters Results

Cluster	C1	C2	C3	C4	C5	C6	E7	E8	E9
0	81.23	11.10	54.00	7.12	42.24	45.19	1.51	16.13	75.16
1	330.47	54.75	233.12	36.86	163.01	165.16	0.07	32.81	181.49
2	964.90	328.80	703.41	234.70	465.36	448.80	0.02	28.76	223.74

Figure 11, depicting the spatial distribution of behavioral mobility clusters, visually reinforces this segmentation, revealing distinct nocturnal activity zones across the urban landscape.

This behavioral provides crucial complementary insights, enabling a richer understanding of nightlife dynamics that incorporate how people move and gather, rather than just where nightlife infrastructure exists.

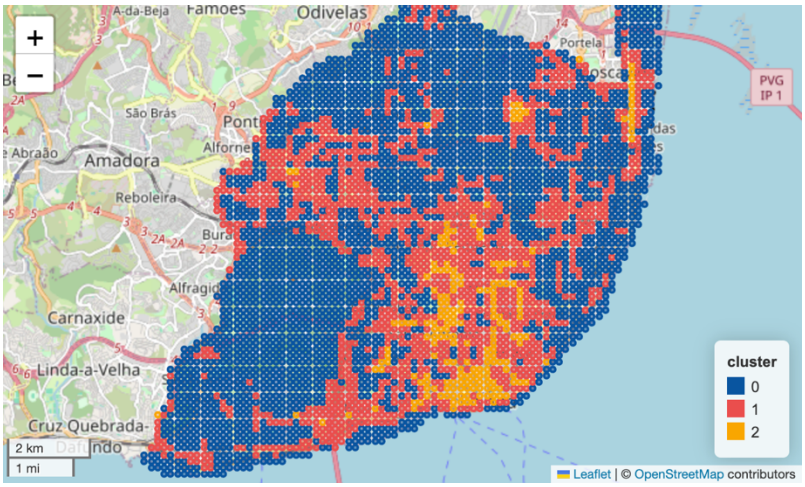


Figure 11 - Spatial distribution of behavioral mobility clusters

3.4.2 Data Treatment

The modelling process considered the 72 grid cells obtained from the intersection between the two nightlife identification approaches, the infrastructure-based method and the behavioral method clustering technique.

Building on these defined spatial units, the model aimed to predict the value of C1 at the subsequent time step ($C1_{t+1}$), which represents the number of distinct mobile terminals detect within each grid cell and serves as a proxy for localized population presence.

To achieve this, a comprehensive feature set was developed to capture temporal dynamics, contextual effects, and meteorological conditions. The resulting features were structured as follows:

Historical variables (lags)

- $C1_{t-1}$, $C1_{t-2}$, C_{t-3}, C_{t-4} – values in the previous 15, 30, 45, 60 minutes, respectively.
- $C1_{t-96}$ – same hour on the previous day.
- $C1_{t-672}$ – same hour on the previous week.

Temporal variables

- Day of week dummies
- Month

Contextual variables

- is_nightlife_period – nightlife period (18:00-05:59).
- is_event_day – occurrence of relevant urban events.

- is_holiday_eve – public holiday eve.

Meteorological variables

- is_raining, is_hot, is_cold, is_windy

A detailed list of the events and holidays, along with their corresponding dates, is presented in Appendix A.

The preprocessing workflow began with the normalization of all numerical variables using the StandardScaler, ensuring zero mean and unit variance to avoid scale-related biases in the models. Categorical variables, namely day_of_week and month, were transformed through One-Hot Encoding with drop_first=True to prevent perfect collinearity. Records containing missing values in any lagged features or in the target variable were removed from the dataset to guarantee complete cases for model training. Finally, the dataset was partitioned into training and testing subsets following an 80/20 split, with random sampling applied to preserve class proportions.

3.4.3 Evaluation Metrics

All models were evaluated based on the following metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2), each offering complementary insights into model performance. Specifically, lower values of error metrics indicate higher predictive accuracy, while R^2 measures the proportion of variance explained by the model, with values closer to 1 representing better fit.

3.4.4 Models and Parameters

This section outlines the configuration of predictive models used to forecast short-term mobility, a supervised regression problem. The chosen algorithms fall into three main categories, progressing from interpretable linear baselines to advanced ensemble and neural network architectures.

The modelling process begins with a Linear Regression model, implemented using the scikit-learn library, serving as a baseline to establish an interpretable reference for the relationships between lagged mobility, temporal, and contextual features. Although limited

in its ability to represent non-linear dynamics, this model provides a transparent benchmark against which more complex algorithms can be evaluated.

Next, three models from the tree-based family were subsequently tested: Decision Tree, Random Forest, and Extreme Gradient Boosting (XGBoost), also implemented using scikit-learn. Decision Trees can model non-linear relationships and interactions without requiring feature scaling. The Random Forest algorithm extends this principle by aggregating multiple decorrelated trees through bagging, reducing variance and improving model stability. Following this principle, XGBoost applies gradient boosting, where trees are sequentially optimized to correct residual errors. This methodological progression from a single tree to bagging and boosting enables a structured evaluation of ensemble strategies within the same family.

Two neural network models were developed: a Multi-Layer Perceptron (MLP) implemented using scikit-learn and a Long Short-Term Memory (LSTM) network implemented through the Keras API with a TensorFlow backend (Keras Team, 2023). The MLP, implemented as a feed-forward neural network, is capable of modeling complex non-linear relationships among mobility, temporal, and contextual variables. Although it does not explicitly account temporal dependencies, it serves as a robust baseline for evaluating non-linear feature interactions within static input windows. Complementarily, the LSTM architecture incorporates sequential dependencies across time steps, enabling the model to learn temporal patterns and dynamic behaviors in mobility flows that evolve throughout the day. Its recurrent structure makes it particularly suitable for capturing time-dependent variations that are critical in mobility forecasting tasks.

To determine the optimal hyperparameter configurations for each algorithm, Randomized Search with 3-fold cross-validation was applied to the non-recurrent ML models. In this procedure, the training data were split into three folds, where in each iteration two folds were used for training and the remaining fold for validation. The method randomly sampled a predefined number of combinations from the specified search spaces (Table 7), allowing efficient exploration without the computational cost of an exhaustive grid search. This approach was particularly suitable given the dataset's size and complexity.

For the LSTM network, however, a distinct tuning strategy was required due to its deep recurrent architecture and high computational cost. Configurations were refined iteratively

based on validation loss monitoring, using early stopping in combination with a ReduceLROnPlateau scheduler to dynamically adjust the learning rate and improve convergence stability. The final dataset comprised approximately 2.4 million temporal sequences, each representing a 4-step (1 hour) sliding window across the 72 grid cells. This volume reflects the total number of valid time windows after applying the defined temporal lag structure. All input features were standardized using a StandardScaler to ensure consistent scaling across grids.

The final configurations for all models are summarized in Table 8.

Table 7 - Hyperparameter search space

Model	Parameters
Linear Regression	No hyperparameters
Decision Tree	max_depth: [5, 10, 15, 20, None] min_samples_split: [2, 5, 10, 20] min_samples_leaf: [1, 2, 4, 8] max_features: [None, 'sqrt', 'log2']
Random Forest	n_estimators: [100, 200, 300, 500] max_depth: [None, 10, 20] min_samples_split: [2, 10] min_samples_leaf: [1, 4] max_features: ['sqrt', 'log2'] bootstrap: [True]
XGBoost	n_estimators: [100, 200, 300, 500] learning_rate: [0.01, 0.05, 0.1, 0.2] max_depth: [3, 4, 5, 6, 7] subsample: [0.6, 0.8, 1.0] colsample_bytree: [0.6, 0.8, 1.0] min_child_weight: [1, 3, 5, 7]

Table 7 - Cont.

MLP	hidden layers: [(50,), (100,), (50,50), (100,50)] activation: [‘relu’,’tanh’] solver: [‘adam’] alpha: [0.0001, 0.01, 0.1] max_iter: [150]

Table 8 - Best hyperparameters

Model	**Parameters**
Linear Regression	No hyperparameters
Decision Tree	min_samples_split: 5; min_samples_leaf: 8; max_features: None; max_depth: 15
Random Forest	n_estimators: 200; min_samples_split: 10; min_samples_leaf: 1; max_features: 'sqrt'; max_depth: 20; bootstrap: True
XGBoost	subsample: 1.0; n_estimators: 500; min_child_weight: 3; max_depth: 7; learning_rate: 0.05; colsample_bytree: 1.0
MLP	hidden layers: (100,50); activation: ‘relu’; solver: ‘adam’; alpha: 0.1; max_iter: 150
LSTM	layers: LSTM (128, 64, tanh) + dense (64, ReLU); dropout: 0.15; optimizer: Adam; learning rate: 0.0003; loss/metric: MSE/MAE; batch size: 128; epochs: 150

CHAPTER 4

Results and Discussion

4.1 Overview

This chapter presents the results of short-term mobility forecasting for the 72 grid cells identified by the intersection of the infrastructure-based and behavioral clustering approaches. The evaluation follows the methodological framework described in Chapter 3, using multiple predictive models and a combination of temporal, contextual, and meteorological features.

Performance is assessed using four complementary metrics to provide a multidimensional view of model accuracy and to better capture temporal and spatial variations in performance, results are reported under multiple perspectives:

1. Global performance – metrics computed across the full 24-hour period, providing an overall measure of model accuracy.
2. Nightlife-focused performance – metrics restricted to the nighttime interval, aligning with the core scope of this research.
3. Per-grid performance – individual metric scores calculated for each of the 72 grid cells, supporting spatial diagnostics and identification of localized model variations.

This multi-level reporting approach allows for the identification of temporal patterns, spatial disparities, and specific areas where predictive accuracy may be systematically lower.

4.2 Model Performance - Global Analysis

The first level of analysis considers global performance across the entire 24-hour period for all models tested. This allows for an overall comparison of accuracy and robustness under general mobility conditions, encompassing both daytime and nighttime dynamics.

Table 9 reports the MSE, MAE, RMSE, and R^2 scores obtained by each model on the test set.

Table 9 - Models results

Model	MSE	MAE	RMSE	R^2
Linear Regression	43171.68	99.19	207.78	0.9072
Decision Tree	38201.95	96.74	195.45	0.9179
Random Forest	32978.60	91.61	181.60	0.9291
XGBoost	36172.65	92.99	190.19	0.9223
MLP	32423.98	92.66	180.07	0.9303
LSTM	34195.51	99.21	184.92	0.9261

The global evaluation reveals differences in performance across the tested models. The linear regression model obtained the weakest results, reflecting its limited capacity to capture the non-linear relationships inherent in mobility dynamics.

Tree-based models, particularly Random Forest and XGBoost, showed improvements, highlighting their ability to model complex feature interactions and reduce variance. Among these, Random Forest achieved the lowest overall error (MSE=32978.60, MAE=91.61, RMSE=181.60) and the highest coefficient of determination (R^2=0.9291), confirming its robustness and generalization capacity under diverse mobility conditions. XGBoost followed closely with slightly higher error metrics but comparable accuracy, reinforcing the reliability of gradient-boosted ensemble methods for short-term mobility forecasting.

The MLP achieved the best overall performance, with an R^2 of 0.9303 and the lowest MSE (32423.98) and RMSE (180.07). This demonstrates the capacity of feed-forward neural networks to learn non-linear temporal and contextual dependencies embedded in the feature space, capturing subtle variations that traditional models may overlook. Despite requiring higher computational effort, the MLP provided the most precise fit to short-term mobility fluctuations, making it the most accurate and globally consistent model.

The LSTM model also delivered competitive performance (R^2=0.9261 and RMSE=184.92), though slightly below that of the top performing approaches. Its recurrent architecture explicitly accounts for temporal dependencies across sequential observations, enabling it to model short-term continuity and dynamic behavioral patterns. While it did not

surpass the MLP or Random Forest in overall accuracy, its temporal sensitivity positions it as a valuable alternative for applications emphasizing real-time prediction and temporal trend analysis.

Overall, the relatively narrow performance gap among the best-performing models (MLP, Random Forest, and LSTM) indicates that multiple algorithmic families can achieve reliable short-term forecasts. Nevertheless, the superior precision of the MLP and temporal learning capability of the LSTM together underscore the growing importance of neural approaches for modeling the complex, dynamic nature of urban mobility.

4.2.1 Best Model

While other models achieved competitive results, the MLP consistently delivered the most robust predictions. Beyond its numerical advantage, the model demonstrates a strong capacity to capture both long-term seasonality and short-term fluctuations, as evidenced in the time series and scatter plots (Figures 12 and 13, respectively). In general, predictions align closely with observed values, with minimal systematic bias and only occasional deviations.

This ability to model complex, non-linear, and context-dependent dynamics justifies its adoption as the benchmark for the forthcoming nightlife-focused (Section 4.3) and per-grid (Section 4.4) analyses, ensuring that subsequent evaluations reflect the most accurate available predictor.

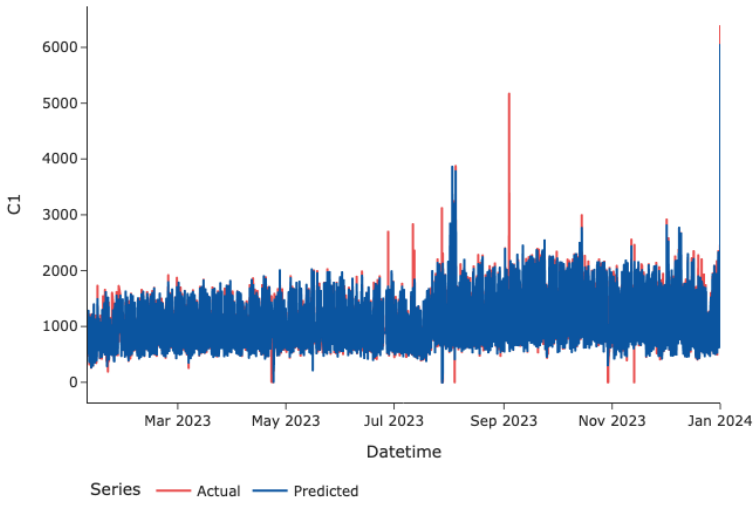


Figure 12 - Time series of actual and predicted values using optimized MLP

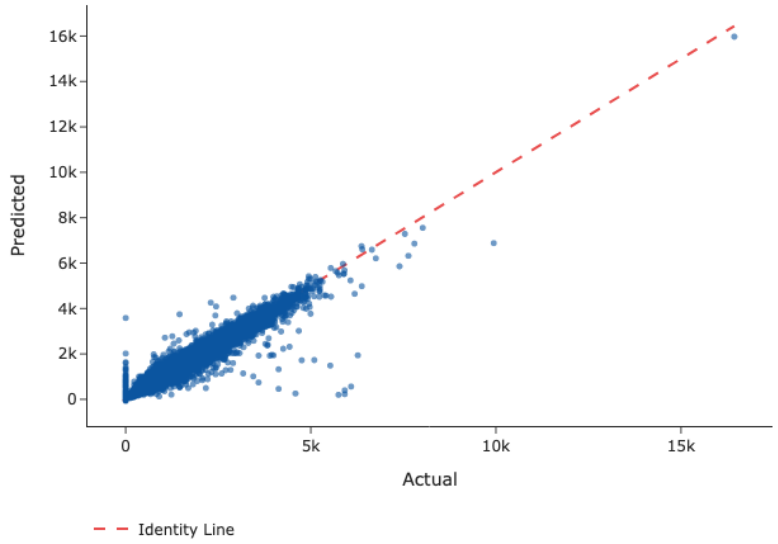


Figure 13 - Scatter plot of actual versus predicted values with the identity line

4.2.2 Permutation-Based Feature Importance Analysis

The permutation-based feature importance analysis provides insights into which factors contributed the most to the performance of the optimized MLP model. Figure 14 shows the top 20 most relevant variables, while Figure 15 presents their aggregated importance by feature group.

The results indicate that the model is dominated by lag features, in other words, past values of the target variable. Specifically, the four most recent lags (C1_previous_15m, C1_previous_30m, C1_previous_45m, and C1_previous_60m) together account for approximately 79% of the total importance, confirming that short-term historical behavior is the primary driver of predictive accuracy. Longer-term dependencies, such as C1_same_hour_previous_day and C1_same_hour_previous_week, add only minor contributions (<3%), reflecting weaker influence of daily or weekly cycles. Temporal indicators (months and day-of-week dummies) account for approximately 15% of the overall importance, highlighting some seasonality effects. Contextual features (e.g., is_nightlife_period, is_event_day) play a marginal role (2%), while meteorological showed no influence, indicating that weather conditions have little to no direct effect on the modeled behavior.

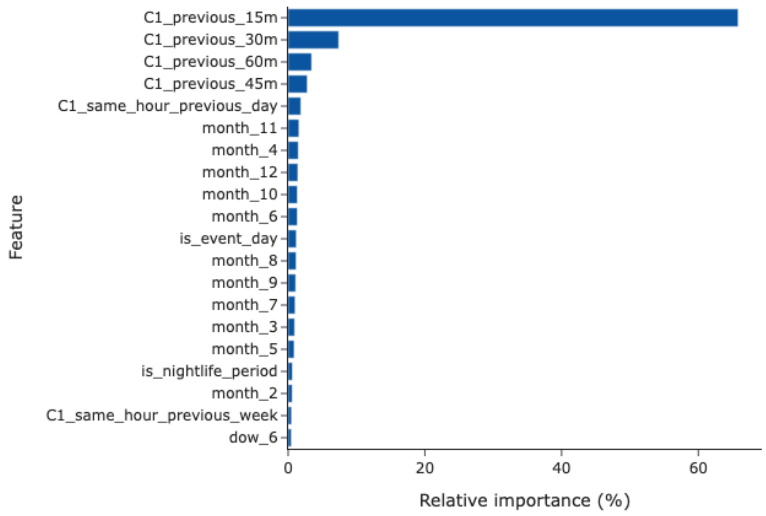


Figure 14 - Top 20 feature importance

The grouped analysis reinforces these findings: lag features dominate explanatory power (82%), with temporal, contextual, and meteorological factors providing only secondary adjustments. Similar conclusions have been reported in previous research, which emphasize that short-term traffic dynamics are largely determined by recent observations, while exogenous variables offer limited additional explanatory power (Barros et al., 2015; Lana et al., 2018). This dominance of lagged variables reflects the strong temporal autocorrelation and inertia inherent in urban mobility systems, where current and immediately preceding conditions are highly predictive of near future states.

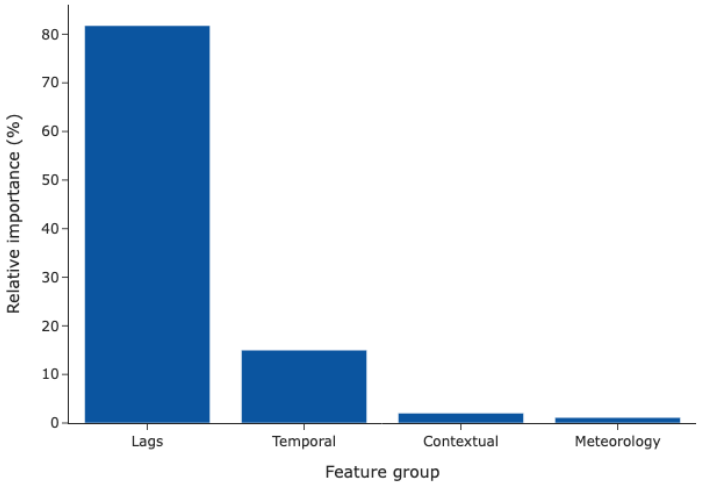


Figure 15 - Feature group importance

Despite this clear dominance, minor contributions from monthly indicators (e.g., April, November, and December) and contextual variables such as is_event_day and is_nightlife_period demonstrate that the model also captures seasonal and social variations in mobility intensity. Furthermore, the presence of the weekday indicator, particularly dow_6 (Sunday), reveals a degree of sensitivity to recurring weekend patterns. These secondary effects, though comparatively small, enhance the interpretability of the model by reflecting nuanced temporal and behavioral influences beyond pure short-term inertia.

4.3 Nightlife - Focused Performance

To align with the primary scope of this research, a second analysis focuses exclusively on the nightlife period between 18:00 and 5:59. As described in Chapter 3, the optimized MLP was trained using all 24-hour data, but with inclusion of a binary feature (is_nightlife_period) that explicitly indicates whether an observation occurs during nightlife hours. This approach allows model to simultaneously learn global mobility dynamics and context-specific patterns associated with nightlife.

For the evaluation, predictions were filtered to retain only nighttime observations in the test set. This enables a direct comparison between overall model performance and its effectiveness in the period of interest.

Table 10 reports the results, showing that while accuracy during nightlife is slightly lower than in the global 24-hour analysis (R^2=0.9225 vs. 0.9303), the model still achieves strong predictive power. The error metrics (MSE=34247.90, MAE=87.61, RMSE=185.06) confirm that the optimized MLP can reliably capture short-term mobility fluctuations at night, despite the greater variability and irregularity typically associated with nocturnal flows.

Table 10 - Performance metrics of the optimized MLP during the nightlife period

Model	MSE	MAE	RMSE	R^2
MLP	34247.90	87.61	185.06	0.9225

4.4 Per-Grid Performance

The third analytical perspective explores spatial variation in model performance by computing metrics separately for each of the 72 grid cells, restricted to the nightlife interval. In this way, this approach allows for a more granular assessment, highlighting areas where the model achieves high predictive accuracy and others where its performance is more limited.

Table 11 reports the descriptive statistics of the main metrics across all grids. On average, the optimized MLP achieved an R^2 of 0.85, confirming strong explanatory power even when restricted to nighttime dynamics. However, the standard deviation of 0.10 reflects heterogeneity between different spatial units. The distribution of error metrics further confirms this variability: while median RMSE is around 146, some grids reach values as high as 757, pointing to localized volatility or irregular demand patterns during nightlife hours.

Table 11 - Summary statistics of per-grid performance metrics

	R^2	MSE	MAE	RMSE	N_test
count	72.00	72.00	72.00	72.00	72.00
mean	0.8535	33560.27	86.94	158.90	3380.50
std	0.1041	68350.10	25.55	91.81	85.68
min	0.2709	6404.65	54.63	80.03	3154.00
25%	0.8326	11051.17	68.22	105.12	3334.25
50%	0.8749	21260.82	80.79	145.81	3389.50
75%	0.9131	33056.89	98.46	181.82	3435.50
max	0.9644	573214.13	189.44	757.11	3547.00

The spatial distribution of predictive accuracy, shown in Figure 16, indicates a generally strong and spatially consistent model performance across Lisbon, with 75% of grids achieving R^2 values above 0.83 and the highest reaching 0.964 (Grid 403). In these areas, the model captures nightlife mobility dynamics with high effectiveness. For example, Grid 455, with an RMSE of 98.7, demonstrates very close agreement between predictive and observed flows, making it one of the most reliable locations. In general, the grids also report relatively

low MAE values (often below 100), indicating not only high explanatory power but also low average prediction error.

Only a few grids show weaker results, reflecting limited explanatory power of the model in those areas. Grid 463 stands out with an R^2 of only 0.27 and an RMSE above 757, suggesting large deviations between predicted and actual values. This location is primarily a daytime commercial area that experiences a sharp decline in activity after dark, previously measured as a 30% drop in nighttime mobility in Section 3.3.4, which limits the model's ability to generalize temporal patterns learned from more stable nightlife areas.

Overall, the spatial pattern of R^2 values confirms that predictive reliability is highest in zones characterized by consistent and high nighttime activity, while deviations occur mainly in areas with transient or irregular movement patterns. This demonstrates the model's robustness under diverse urban conditions, with localized declines linked to functional and behavioral variability.

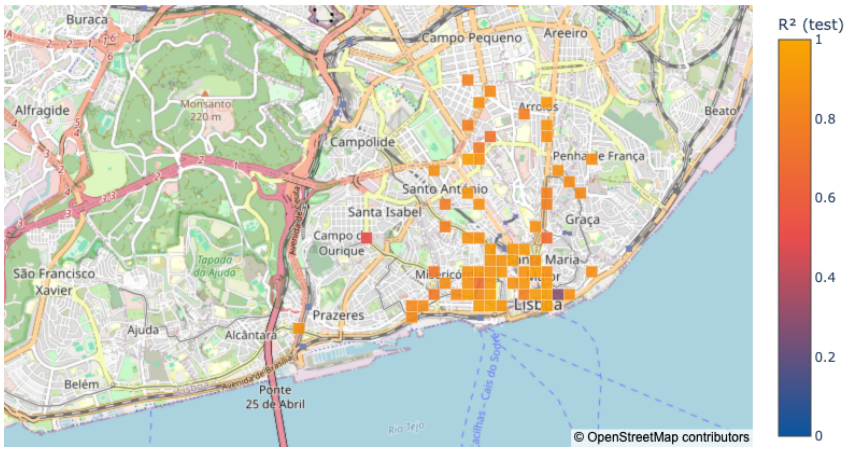


Figure 16 - Spatial distribution of model performance (R^2)

CHAPTER 5

Conclusion

5.1 Main Findings and Conclusions

This research explores how mobile network data can be leveraged to analyze and predict nighttime human concentration in urban nightlife areas. By integrating spatial analysis, behavioral clustering, and predictive modeling, the study demonstrated the value of large-scale mobility data as tool for understanding and anticipating nocturnal urban dynamics.

Addressing the first objective (O1), discussed in Chapter 2, the review of the existing literature revealed a clear evolution from static and descriptive analyses of urban mobility toward dynamic, predictive, and data-driven approaches. The incorporation of mobile phone data, combined with the advances in ML, has transformed how urban phenomena are measured and modeled. Recent studies have increasingly relied on supervised learning methods such as Random Forests, Gradient Boosting, and Neural Networks for mobility prediction, alongside unsupervised approaches such as clustering analysis to uncover latent spatial and behavioral structures. In line with the methodological trends, the present research adopted a hybrid framework that integrates unsupervised clustering for identifying nightlife activity zones and supervised learning models for short term population prediction. Nevertheless, despite this methodological evolution, the literature still reveals persistent gaps, particularly regarding nighttime mobility, ethical data use and governance, and the limited incorporation of behavioral insights into urban planning frameworks, an area that this research sought to explore.

Concerning the second objective (O2), developed in Section 3.3.2 and Section 3.4.1, the analysis identified and spatially characterized nightlife zones by the combination of geolocated POIs and mobile network grid data. Using spatial joins between nightlife POIs and grid geometries, a total of 227 nightlife grids were detected concentrated mainly in historically vibrant areas such as Cais do Sodré, Bairro Alto, Santos and Baixa. The results revealed that POI density alone does not fully capture actual nighttime dynamics, as several grids with limited nightlife infrastructure exhibited high mobility levels, while others with numerous venues showed moderate activity. This mismatch underscores the need to complement static infrastructure-based approaches with behavioral evidence derived from

mobility data. To overcome this limitation, the K-Means algorithm was applied, identifying three clusters, one of which comprising 199 grids emerged as the most active and persistent during nighttime hours.

The modeling process considered the 72 grid cells obtained from the intersection between the two nightlife identification approaches. This intersection ensured that only areas consistently identified as nightlife active, both structurally and behaviorally, were retained for predicting modeling.

Regarding the third objective (O3), addressed in Chapter 4, the study developed and evaluated several supervised learning models to forecast short-term variations in population presence within these 72 selected nightlife grids. Among the tested models, the MLP achieved the highest predictive accuracy (R^2=0.93), followed closely by Random Forest and LSTM architectures. These results highlight the potential of neural networks and ensemble approaches to capture both spatial and temporal dependencies inherent to human mobility behavior. Feature importance analysis showed that model performance was mainly driven by short-term lag variables, indicating strong temporal continuity in mobility behavior, while temporal and contextual factors contributed marginally, and meteorological conditions had no significant effect.

In fulfilling the fourth objective (O4), which builds on insights from the exploratory data analysis in Section 3.3.4 and the empirical findings discussed in Chapter 4, the research provided actionable insights that enhance decision support for data-driven urban management and planning. The spatial and temporal analysis revealed that while overall mobility decreases after dark, specific areas sustain intense nighttime activity, particularly during weekends. These findings confirm their continuous role as social and economic attractors within the city, while areas with predominantly daytime commercial functions showed sharp declines after working hours, illustrating functional contrasts across the urban fabric. By integrating these insights with predictive modeling framework, this study demonstrates how mobility analytics can support evidence-based decision making in multiple domains. The proposed models can inform tactical decisions, such as optimizing public transport supply, allocating crowd control resources, or adjusting safety measures in real time, as well as strategic planning and policy design decisions related to zoning, noise mitigation, and the sustainable regulation of nighttime economies.

5.2 Limitations

While the findings demonstrate the potential of mobile network data for analyzing and forecasting nightlife dynamics, several limitations must be acknowledged.

The use of mobile terminal (C1) counts as a proxy for population presence does not distinguish between residents, workers, or tourists, limiting the social interpretation of mobility patterns. Moreover, the spatial (200m) and temporal (15 min) resolution, while suitable for city-scale analysis, may smooth out localized variations in movement and crowd density. Additionally, the limited granularity of meteorological and event level information further constrains the explanatory capacity of exogenous features.

These limitations stem mainly from the datasets used. However, additional methodological limitations were identified in the hyperparameter optimization process. For the non-recurrent models, Randomized Search offered an efficient solution but explored only part of the parameter space, which may overlook optimal configurations. For the LSTM network, tuning was performed iteratively through manual adjustments based on validation loss, which, while improving convergence, limited automation and reproducibility.

5.3 Future Work

Future work should address these limitations in several ways. From a data perspective, integrating multimodal data sources, including public transport usage, social media activity, or high-resolution environmental data and event datasets, would enrich the contextual understanding of nighttime mobility. Extending the proposed framework to different urban contexts (e.g., other cities) would also enable broader evaluation of its generalizability and operational relevance.

Methodologically, adopting more systematic and automated hyperparameter optimization frameworks, such as Bayesian Optimization or Optuna-based searches, could improve efficient and reproducible exploration of the parameter space. Furthermore, exploring advanced spatiotemporal architectures such as Conv-LSTM, has the potential to better capture local spatial dependencies and short-term temporal dynamics of urban mobility behavior.

To conclude, embedding nighttime mobility forecasting tools into urban decision-support systems could enhance evidence-based planning for sustainable, inclusive, and safe nighttime economies, strengthening the link between data analytics and practical urban governance.

References

Alam, N., & O'Connell, D. (2024). Nightlife human behavior. *Journal of Human Behavior in the Social Environment*, 1–13. https://doi.org/10.1080/10911359.2024.2446824

Ali, A., Zhu, Y., & Zakarya, M. (2021). A data aggregation based approach to exploit dynamic spatio-temporal correlations for citywide crowd flows prediction in fog computing. *Multimedia Tools and Applications*, *80*(20), 31401–31433. https://doi.org/10.1007/s11042-020-10486-4

Bao, W., Gong, A., Zhang, T., Zhao, Y., Li, B., & Chen, S. (2023). Mapping Population Distribution with High Spatiotemporal Resolution in Beijing Using Baidu Heat Map Data. *Remote Sensing*, *15*(2), 458. https://doi.org/10.3390/rs15020458

Barros, J., Araujo, M., & Rossetti, R. J. F. (2015). Short-term real-time traffic prediction methods: A survey. *2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, 132–139. https://doi.org/10.1109/MTITS.2015.7223248

Bellodi, E., Zese, R., Petrovich, C., Frascella, A., & Bertasi, F. (2024). Predicting the impact of public events and mobility in Smart Cities. *IET Smart Cities*, *6*(4), 253–275. https://doi.org/10.1049/smc2.12087

Berlotti, M., Di Grande, S., & Cavalieri, S. (2024). Proposal of a Machine Learning Approach for Traffic Flow Prediction. *Sensors*, *24*(7), 2348. https://doi.org/10.3390/s24072348

Bernini, A., Toure, A. L., & Casagrandi, R. (2019). The time varying network of urban space uses in Milan. *Applied Network Science*, *4*(1), 128. https://doi.org/10.1007/s41109-019-0245-x

Bertini, F., Davtian, T., & Sharma, R. (2023). Understanding cycling mobility: Bologna case study. *Computational Urban Science*, *3*(1), 3. https://doi.org/10.1007/s43762-022-00073-8

Bjerre-Nielsen, A., Minor, K., Sapieżyński, P., Lehmann, S., & Lassen, D. D. (2020). Inferring transportation mode from smartphone sensors: Evaluating the potential of Wi-Fi and Bluetooth. *PLOS ONE*, *15*(7), e0234003. https://doi.org/10.1371/journal.pone.0234003

Böcker, L., Dijst, M., & Prillwitz, J. (2013). Impact of Everyday Weather on Individual Daily Travel Behaviours in Perspective: A Literature Review. *Transport Reviews*, *33*(1), 71–91. https://doi.org/10.1080/01441647.2012.747114

Câmara Municipal de Lisboa. (2025). *Apresentação – LxDataLab*. https://lisboaaberta.cm-lisboa.pt/index.php/pt/lx-data-lab/apresentacao

Christidis, P., Gonzalo, M. V., & Radics, M. (2023). *Modelling daily mobility using mobile data traffic at fine spatiotemporal scale* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2311.09683

Crivellari, A., & Resch, B. (2022). Investigating functional consistency of mobility-related urban zones via motion-driven embedding vectors and local POI-type distributions. *Computational Urban Science*, *2*(1), 19. https://doi.org/ 10.1007/s43762-022-00049-8

Ferreira, J. C., Francisco, B., Elvas, L., Nunes, M., & Afonso, J. A. (2023). Predicting People's Concentration and Movements in a Smart City. *Electronics*, *13*(1), 96. https://doi.org/10.3390/electronics13010096

Gariazzo, C., Pelliccioni, A., & Bogliolo, M. P. (2019). Spatiotemporal Analysis of Urban Mobility Using Aggregate Mobile Phone Derived Presence and Demographic Data: A Case Study in the City of Rome, Italy. *Data*, *4*(1), 8. https://doi.org/10.3390/data4010008

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

Haddaway, N. R., Page, M. J., Pritchard, C. C., & McGuinness, L. A. (2022). *PRISMA2020*: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. *Campbell Systematic Reviews*, *18*(2), e1230. https://doi.org/10.1002/cl2.1230

Hashemi, M., & Zufle, A. (2025). *From Points to Places: Towards Human Mobility-Driven Spatiotemporal Foundation Models via Understanding Places* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2506.14570

Hong, Y., Zhang, Y., Schindler, K., & Raubal, M. (2025). *Deep Generative Model for Human Mobility Behavior* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2510.06473

Kaluarachchi, Y. (2022). Implementing Data-Driven Smart City Applications for Future Cities. *Smart Cities*, *5*(2), 455–474. https://doi.org/10.3390/smartcities5020025

Keras Team. (2023). *Keras 3: A New Multi-Backend Keras* (Version 3) [Computer software]. GitHub repository. https://github.com/keras-team/keras

Kim, Y.-L. (2020). Data-driven approach to characterize urban vitality: How spatiotemporal context dynamically defines Seoul's nighttime. *International Journal of Geographical Information Science*, *34*(6), 1235–1256. https://doi.org/10.1080/13658816.2019.1694680

Kumar, H., Singh, M. K., Gupta, M. P., & Madaan, J. (2020). Moving towards smart cities: Solutions that lead to the Smart City Transformation Framework. *Technological Forecasting and Social Change*, *153*, 119281. https://doi.org/10.1016/j.techfore.2018.04.024

Lana, I., Del Ser, J., Velez, M., & Vlahogianni, E. I. (2018). Road Traffic Forecasting: Recent Advances and New Challenges. *IEEE Intelligent Transportation Systems Magazine*, *10*(2), 93–109. https://doi.org/10.1109/MITS.2018.2806634

Lee, C., & Lee, E. H. (2024). Evaluation of urban nightlife attractiveness for Millennials and Generation Z. *Cities*, *149*, 104934. https://doi.org/10.1016/j.cities.2024.104934

Lin, Y., Shen, Z., & Teng, X. (2021). Review on Data Sharing in Smart City Planning Based on Mobile Phone Signaling Big Data: From the Perspective of China Experience: Anonymization VS De-anonymization. *International Review for Spatial Planning and Sustainable Development*, *9*(2), 76–93. https://doi.org/10.14246/irspsd.9.2_76

Liu, Y., & Dong, B. (2024). Modeling urban scale human mobility through big data analysis and machine learning. *Building Simulation*, *17*(1), 3–21. https://doi.org/10.1007/s12273-023-1043-z

Open Geospatial Consortium. (2011). *OpenGIS Implementation Specification for Geographic Information – Simple Feature Access – Part 1: Common Architecture*. https://www.ogc.org/standards/sfa

Open-Meteo. (2025). *Open-Meteo Archive API: Free historical weather data API*. https://open-meteo.com/

OpenStreetMap contributors. (2025). *OpenStreetMap*. https://www.openstreetmap.org/

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, n71. https://doi.org/10.1136/bmj.n71

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Qi, G., Huang, A., Guan, W., & Fan, L. (2019). Analysis and Prediction of Regional Mobility Patterns of Bus Travellers Using Smart Card Data and Points of Interest Data. *IEEE Transactions on Intelligent Transportation Systems*, *20*(4), 1197–1214. https://doi.org/10.1109/TITS.2018.2840122

Reia, J. (2025). Invisible data in night-time governance: Addressing policy gaps and building a digital rights framework for cities after dark. *Data & Policy*, *7*, e17. https://doi.org/10.1017/dap.2024.90

Roy, S. S., Perlman, J., & Balling, R. C. (2020). Analysis of urban mobility in South Florida using Uber Movement. *Case Studies on Transport Policy*, *8*(4), 1393–1400. https://doi.org/10.1016/j.cstp.2020.10.003

Santa, F., Henriques, R., Torres-Sospedra, J., & Pebesma, E. (2019). A Statistical Approach for Studying the Spatio-Temporal Distribution of Geolocated Tweets in Urban Environments. *Sustainability*, *11*(3), 595. https://doi.org/10.3390/su11030595

Santiago-Iglesias, E., Romanillos, G., Carpio-Pinedo, J., Sun, W., & García-Palomares, J. C. (2024). Recovering urban nightlife: COVID-19 insights from Google Places activity trends in Madrid. *Journal of Maps*, *20*(1), 2371927. https://doi.org/10.1080/17445647.2024.2371927

Shi, M. J., Cao, Q., Van Rompaey, A., Pu, M., & Ran, B. (2023). Modeling vibrant areas at nighttime: A machine learning-based analytical framework for urban regeneration. *Sustainable Cities and Society*, *99*, 104920. https://doi.org/10.1016/j.scs.2023.104920

Shi, P., Xiao, Y., & Zhan, Q. (2020). A Study on Spatial and Temporal Aggregation Patterns of Urban Population in Wuhan City based on Baidu Heat Map and POI Data. *International Review for Spatial Planning and Sustainable Development*, *8*(3), 101–121. https://doi.org/10.14246/irspsda.8.3_101

Son, N. N., Thu, N. T. P., Dung, N. Q., Huyen, B. T. T., & Xuan, V. N. (2023). Determinants of the Sustained Development of the Night-Time Economy: The Case of Hanoi, Capital of Vietnam. *Journal of Risk and Financial Management*, *16*(8), 351. https://doi.org/10.3390/jrfm16080351

Wang, D., Chen, H., Li, C., & Liu, E. (2023). Exploring the Relationship between Land Use and Congestion Source in Xi'an: A Multisource Data Analysis Approach. *Sustainability*, *15*(12), 9328. https://doi.org/10.3390/su15129328

Wasserman, J., Gerard, J., Tratner, J., Perry, M., Badaracco, A. G., Farmer, C., Hjelle, G. A., Snow, A. D., Cochran, M., Gillies, S., Culbertson, L., Bartos, M., Eubank, N., Maxalbert, Bilogur, A., Rey, S., Ren, C., Arribas-Bel, D., Wasser, L., … Leblanc, F. (2020). *geopandas/geopandas: V0.8.1* (Version v0.8.1) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.3946761

Wei, B., Su, G., & Liu, F. (2023). Dynamic Assessment of Spatiotemporal Population Distribution Based on Mobile Phone Data: A Case Study in Xining City, China. *International Journal of Disaster Risk Science*, *14*(4), 649–665. https://doi.org/10.1007/s13753-023-00480-3

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. *In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, *1*, 29–39.

Wu, Y., Wang, L., Fan, L., Yang, M., Zhang, Y., & Feng, Y. (2020). Comparison of the spatiotemporal mobility patterns among typical subgroups of the actual population with mobile phone data: A case study of Beijing. *Cities*, *100*, 102670. https://doi.org/10.1016/j.cities.2020.102670

Xu, Y., Li, J., Xue, J., Park, S., & Li, Q. (2021). Tourism Geography through the Lens of Time Use: A Computational Framework Using Fine-Grained Mobile Phone Data. *Annals of the American Association of Geographers*, *111*(5), 1420–1444. https://doi.org/10.1080/24694452.2020.1812372

Zhang, Y., Liu, J., Wang, Y., Cao, Y., & Bai, Y. (2021). Research on the Method of Urban Jobs-Housing Space Recognition Combining Trajectory and POI Data. *ISPRS International Journal of Geo-Information*, *10*(2), 71. https://doi.org/10.3390/ijgi10020071

Zhao, G., & Yang, M. (2020). Urban Population Distribution Mapping with Multisource Geospatial Data Based on Zonal Strategy. *ISPRS International Journal of Geo-Information*, *9*(11), 654. https://doi.org/10.3390/ijgi9110654

Zufiria, P. J., Pastor-Escuredo, D., Úbeda-Medina, L., Hernandez-Medina, M. A., Barriales-Valbuena, I., Morales, A. J., Jacques, D. C., Nkwambi, W., Diop, M. B., Quinn, J., Hidalgo-Sanchís, P., & Luengo-Oroz, M. (2018). Identifying seasonal mobility profiles from anonymized and aggregated mobile phone data. Application in food security. *PLOS ONE*, *13*(4), e0195714. https://doi.org/10.1371/journal.pone.0195714

Appendix A – List of events

Event	Date
Santos Populares	June 1-30
Lisbon Pride	June 24-25
World Youth Day	August 1-6
Web Summit	November 13-16
Lisboa Games Week	November 23-26

Holiday	Date
New Year's Day	January 1
Good Friday	April 7
Easter Sunday	April 9
Liberty Day	April 25
Labour Day	May 1
Corpus Christi	June 8
Portugal Day	June 10
St. Anthony's Day (Lisbon)	June 13
Assumption of Mary	August 15
Implantation of the Republic	October 5
All Saints' Day	November 1
Restoration of Independence	December 1
Immaculate Conception	December 8
Christmas Day	December 25