INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Analysis of Commuter Movements on Main Access Routes to the City of Lisbon Using Cell Phone Data

Ricardo Nuno Costa Moinhos Santos de Sousa

Master's degree in integrated business Intelligence Systems

Supervisor
PHD Joao Carlos Amaro Ferreira, Assistant Professor with Aggregation
ISCTE-University Institute of Lisbon

Co-Supervisor
MSc Bruno Alexandre Mateus Francisco, Invited Senior Assistant
ISCTE-University Institute of Lisbon

May, 2024

TECNOLOGIAS
E ARQUITETURA

Analysis of Commuter Movements on Main Access Routes to the City of Lisbon Using Cell Phone Data

Ricardo Nuno Costa Moinhos Santos de Sousa

Master's degree in integrated business Intelligence Systems

Supervisor
PHD Joao Carlos Amaro Ferreira, Assistant Professor with Aggregation
ISCTE-University Institute of Lisbon

Co-Supervisor
MSc Bruno Alexandre Mateus Francisco, Invited Senior Assistant
ISCTE-University Institute of Lisbon

May, 2024

To my beloved Parents, Grandmother and Family

Acknowledgements

This adventure began in 2021 when I was finishing my bachelor's degree and decided that I would complete my studies with a master's degree outside the area I had initially chosen (Management). That's when the opportunity arose to join MSIAD, to educate myself in the field of IT with a focus on decision-making, which I thought would round off my education nicely. Along the way, I had several teachers who helped me and provided enough resources to complete my studies. I would like to thank Professor João Ferreira and Professor Bruno Francisco, who were always helpful in solving the problems I encountered, for agreeing to guide me through this thesis.

I would also like to thank my family and the people closest to me for helping me through school and my life.

Resumo

Este trabalho aborda o problema do congestionamento nas principais vias de acesso à cidade de Lisboa, e os principais fatores que estão na sua origem. Foi realizado com a ajuda da Câmara Municipal de Lisboa, via Lx Data Lab, que disponibilizou os dados utilizados através de um contrato já estabelecido com uma operadora móvel e com o IPMA, sendo possível gerar eventos anonimamente através da sinalização trocada entre a rede móvel e o terminal do utilizador, obtendo informação sobre o volume de entradas e saídas a cada 5 minutos nos 11 pontos de acesso à cidade durante o período de setembro de 2021 a dezembro de 2022. Disponibilizaram também acesso às condições climatéricas e pluviométricas durante o período de análise, permitindo comparar estas condições com o volume de entradas e saídas, caracterizando-as nos períodos da hora de ponta da manhã e da tarde e com variáveis como os períodos escolares e de férias.

Este trabalho teve como objetivos: 1) criar conhecimento sobre o problema do congestionamento e da gestão do trânsito; 2) identificar os possíveis fatores e períodos horários com mais trânsito; 3) visualizar e analisar o trabalho realizado de forma a extrair conhecimento que auxilie a tomada de decisão na resolução do problema do congestionamento. Para a realização de previsões de curto prazo foi utilizado um modelo LTSM, obtendo bons resultados, quer pela sua capacidade de prever os padrões de subida e descida dos valores reais, quer pelo baixo valor de perda no conjunto de validação.

Palavras-chave: Congestionamento, Cidades Inteligentes, Rede Celular, Roaming, Acesso à Cidade e Análise de Dados.

Abstract

This work aims to tackle the problem of congestion on the main access roads to the city of Lisbon, and the main factors behind it. With the help of the Lisbon City Council, via Lx Data Lab, which provided the date obtained from a contract already established with a mobile operator and with IPMA, where it is possible to generate events completely anonymously through the signalling exchanged between the mobile network and the user's terminal to obtain information on the volume of entries and exits every 5 minutes at the 11 access points to the city during the period from September 2021 to December 2022. They also provided access to the weather and rainfall conditions during the same period of analysis, allowing a comparison to be made between these conditions and the volume of inflows and outflows, characterising them during the morning and afternoon rush hour periods and with variables such as school and holiday periods.

This work aimed to: 1) create knowledge about the problems of congestion and traffic management; 2) identify the possible factors and time periods with the most traffic; 3) visualise and analyse the work carried out in order to extract knowledge that will help decision-making in tackling the problem of congestion. An LTSM model was used to make short-term forecasts, which obtained good results both in terms of its ability to predict the patterns of rise and fall of the real values and in obtaining a low loss value in the validation set.

Keywords: Congestion, Smart Cities, Cellular Network, Roaming, City Access and Data Analysis.

Index

Acknowledgements iii

Resumo vii

Abstract ix

Chapter 1 1

Introduction 1

1.1. Introduction 1

1.2. Framework and Motivation 1

1.3. Objectives 2

1.4. Work Challenges 3

1.5. Work Outline 3

Chapter 2 5

Literature Review 5

2.1. Related work 5

2.2. Search and inclusion tecniques 9

Chapter 3 11

Knowledge Extraction Approach 11

3.1. Business Understanding 11

3.2. Data description and understanding 11

3.3. Data preparation 13

3.4. Visualization 14

Chapter 4 15

Visualizations 15

4.1. Insights and visualizations 15

4.2. What are the busiest times of day? 15

4.3. Which days of the week have the most traffic? 16

4.4. Which months have the most traffic? 17

4.5. What are the busiest school periods? 19

4.6. How does rainfall influence traffic? 21

Chapter 5 25

Traffic Prediction and Detection 25

5.1. Predictions 25

Chapter 6 31

Conclusions and Future Work 31

6.1. Conclusions 31

6.2. Future work 32

References 33

List of Figures

Figure 1- Evolution of relevant studies per year 10
Figure 2- Average evolution of commuter movements throughout the day..... 15
Figure 3- Average commuter movements by day of the week (morning rush hour) 16
Figure 4- Average commuter movements by day of the week (afternoon rush hour)..... 17
Figure 5- Average commuter movements by month (morning rush hour)..... 18
Figure 6- Average commuter movements by month (afternoon rush hour)..... 19
Figure 7- Average commuter movements by school term (morning rush hour) 20
Figure 8- Average commuter movements by school term (afternoon rush hour) 20
Figure 9- Average commuter movements by rainfall level..... 21
Figure 10- Average commuter movements by precacumulada values (morning rush hour period) 22
Figure 11- Average commuter movements by precacumulada values (afternoon rush hour period).... 23
Figure 12- Trafic prediction in rush hour periods in the 11 Lisbon entry points 26
Figure 13- Trafic prediction in the morning rush hour periods in the 11 Lisbon entry points..... 27
Figure 14- Trafic prediction in the afternoon rush hour periods in the 11 Lisbon entry points..... 28

List of Tables

Table 1- Keywords definition..... 10
Table 2- Commuter movements dataset variables..... 13
Table 3- Meteorology dataset variables 13
Table 4- Minimum loss and val_loss values shown for peak hour periods, for 50 epochs 26
Table 5- Minimum loss and val_loss values shown for the morning rush hour period, for 50 epochs . 27
Table 6- Minimum loss and val_loss values shown for the afternoon rush hour period, for 50 epochs 28

CHAPTER 1

Introduction

1.1. Introduction

The characterization of traffic is essential for the planning of life in a city like Lisbon, particularly regarding the volume of people who circulate there daily, during rush hour periods, causing congestion on the main access roads to the city [1].

Smart cities appear to face these and other urban challenges, using technology and constant innovation to promote sustainable development and improve the quality of life in cities [2], [3], [4]. Through the use of data, we are able to plan and forecast traffic at major access points [5], [6], [7], that helps us understand what affects the traffic in Lisbon and how it varies over time.

1.2. Framework and Motivation

According to [8], [9], around 380 thousand people enter the city of Lisbon every day and 47 thousand leave it. As a result of these numbers, every day the city of Lisbon sees its number of users increase by more than 70%, as a result of commuting movements from home to work and home to school [10]. Since the city of Lisbon has a population of over 500,000 and an urban area of around 100 km2, these commuter movements cause greater congestion and negatively affect urban mobility, causing delays in commuting. The time spent commuting increases stress and has a negative impact on the mental and physical health of individuals, reducing their quality of life. This situation is further propagated with the environmental implications that congestion causes, with a greater emission of pollutants.

This work was developed with the help of the Lisbon City Council [11], via Lx Data Lab[12], which provided the date obtained from a contract already established with a mobile operator and with IPMA [13].

According to the [4], the use of IoT technologies and data analytics increases quality of life by optimizing the efficiency of city operations and services, so a coordinated multi-modal public transport system is needed to face extreme traffic congestions created by the growth of big cities.

A smart city can be implemented to optimize all the operations of a city at different levels to combat constant congestion at the main entry points to the city. This level of efficiency, we believe, can only be achieved when we use data to better plan traffic systems that cause delays and poorer quality of life and the environment, identifying the hours of greatest traffic congestion, its causes and how to make decisions to mitigate them, based on the information obtained from the analysis of the results and forecasts of historical data. By identifying major congestion points, it will be possible to understand whether the optimization of services provided by the city is sufficient to counteract traffic at the main entry and exit points at peak times, or even give an idea whether it will be necessary to improve existing access roads to the city or whether there is a need to create new roads to achieve greater traffic efficiency.

Having this in mind, my motivation to develop this academic work came from the challenge of Data Lab which aimed to study the number of mobile devices circulating the main access points to the city of Lisbon, how the flows of these numbers vary and what affects them. The possibility to transform this data in practical applicability, with the intention to provide decision makers with valid information which can be used to tackle the problems of congestion that affect the quality of life of individuals and the environment, was a selling point for this work.

1.3. Objectives

This study aims to analyze and visualize the date of the number of mobile devices entering and exiting through each of the 11 entry and exit points in the Lisbon City Hall, with the intend of creating knowledge in relation to:

1. Characterize the volume of entries and exits during the morning rush hour period (7:30 AM-10:00 AM).
2. Characterize the volume of entries and exits during the afternoon rush hour period (5:00 PM-7:30 PM).
3. Relate the two previous points with variables such as school or vacation periods and the existence of rainfall.
4. Variation of the volume of entries and exits by periods of the day and entry and exit points.

The analysis will focus on the 16-month period from September 2021 to the end of 2022, mostly through visualizations, to gain a greater understanding of the patterns of

inflow and outflow volumes at the 11 identified points, and how these patterns are affected by the weather and the differences over the period. This work uses CRISP-DM data science methodology, using past data from mobile data, that uses cellular grid areas with information about the volume of entries and exists in a time stamp period of five minutes.

1.4. Work Challenges

In the making of this work, these are the major challenges that we face:

1. Data volume and complexity: In this work we had to join two data-rich databases with different information that had to be carefully worked and filtered according to our objectives, which presented significant challenges.
2. Data cleaning and pre-processing: Due to working with extensive databases, to ensure quality and reliability of the presented results, it was necessary to clean and pre-process the data, through the identification and treatment of missing values, inconsistencies and errors.
3. Data enrichment: Additional information like weekends, holidays, school calendar was essential for our project, this information had to be correctly added to the database to meet the results outlined for this work.

1.5. Work Outline

This work is divided into 5 chapters including this first introductory chapter which is concluded. Thus, the structure of this work is as follows:

- Chapter 2: The techniques used in the literature review are described, as well as topics and projects already developed in similar areas that allow to create knowledge.
- Chapter 3: Describes and shows the description of the methodology and date used, as well as its preparation and the context of the visualizations developed.
- Chapter 4: This chapter responds to the objectives established to create knowledge in this work, making a detailed analysis of what is intended to be validated.
- Chapter 5: Demonstrates the tests of traffic forecasting and detection carried out in this study, using LSTM models.
- Chapter 6: Includes the conclusions reached with this work and the possible developments that can be made to it.

CHAPTER 2

Literature Review

2.1. Related work

According to ANACOM, in Portugal there are around 13.5 million active sim cards, which gives us an opportunity to have a large number of mobile phones at our disposal, that are constantly emitting signals of their movements, allowing us to have enough information to trace their patterns and behaviours. This pool of great information can encourage various studies in areas involving the movement of individuals, as it can serve as an aid in the research work carried out by different researchers.

The article [1] identifies traffic congestion as a major problem in most cities affecting it at different levels, stating that they can arise quickly but can also extend over a long period of time, causing traffic density to exceed a bottleneck density threshold. When observed hour by hour, specific characteristics of sudden traffic jams were detected by the study, which also concluded that chaotic driving patterns and inadequate traffic management result in intense and more frequent congestion, and poor utilisation of road capacity. The study thus concludes that tackling the problem of congestion in cities requires smart management of road infrastructure, not just the construction of new ones.

The development and reorganisation of road networks in large cities has led to challenges in controlling traffic congestion. The [14] study investigated the relationship between urban space use patterns in road areas and traffic congestion in the new cities of Xi'an, China, by using a DBSCAN algorithm that clusters and analyses POIs (points of interest). They found that POIs with bi-directional flows of entrances and exits are more likely to cause road congestion than those with unidirectional flows, and that POIs with flexible traffic as opposed to those with predicted traffic have a statistical correlation with increased congestion. The results of the study help urban planners to tackle congestion by optimising the socio-economic functions of roads.

In the article [3], based on India's congestion problem, they propose an IoT system focussed on maintaining congestion, clearing emergency vehicles and detecting stolen vehicles, to solve or mitigate these problems. The system used in this paper is composed of four RF module setups controlled by a hub RF module setup wirelessly that aims to reduce

manpower and costs of traffic control to achieve an efficient maintenance system, thus this work will allow to improve the safe flow of traffic and reduce chaos in its management.

The work, [2] explored the use of traffic congestion detection (TCD) techniques to solve traffic jam problems. In this paper they used two algorithms, a TCD algorithm to detect congestion based on traffic features, which starts by cleaning and pre-processing the data to then calculate the absolute value of the first derivative of each sample in order to detect anomalies, calculate its probability and classify it; and an EB-TCD algorithm that uses information contained in several traffic features in parallel, proposing a technique that combines the anomaly scores of the different traffic features into a single score in order to mitigate the effects of noise in traffic data by rapidly detecting changes in traffic patterns. The results of this study thus demonstrated that both algorithms outperformed other commonly used algorithms in detection time and false alarm rates while maintaining a detection rate at a similar high level.

In the paper, [15] was another paper that addressed congestion problems considering smart cities concepts, having managed to implement a short-term traffic prediction model that guarantees a better unloading performance as well as a high prediction quality. This model was based on EC and deep learning that analysed the performance by simulations, resulting in an algorithm with a lower average delay cost, maintaining an average unloading utility of 70%, but obtaining a recognition accuracy (98.06%) at least 1.14% higher than that of the advanced convolution neural network (CNN) algorithm proposed by other academics, achieving a faster convergence rate and a better prediction performance. Thus, the results of this work want to provide a good experimental basis for traffic flow forecasting and a better implementation of smart cities.

There is a study [4] that addresses the need to implement a system using IoT technologies and data to improve the efficiency of public transport that are already facing challenges in maintaining services in large cities, to combat congestion problems by providing a better supply and improvement of mobility that improves the quality of life. This congestion, affected by the growth of large cities in the centre and urban sprawl in the peripheries, together with environmental concerns and the recent pandemic, has resulted in changes in work and housing patterns, which alter the movements of individuals throughout the city. This article thus highlights the need for smart cities to have public transport that offers ubiquitous access with real-time response to demand, and addresses the challenges

for creating, planning and managing a smart public transport system, also presenting a survey of recent research and innovations in the field under study.

The scientific article [16] mentions the increased use of data in previous years, entering an era of big data in transport, however the forecasting methods used were still very shallow and without much real application. Thus, the paper proposes the use of an original deep learning prediction method that considers intrinsic temporal and spatial correlations that utilises autoencoders as building blocks to represent the traits of traffic flows for prediction. The results of these experiments show that the proposed prediction method has superior performance.

There are other studies that speak to the inefficiency in traffic flow prediction of previously used systems, using different deep learning models. The first one is the [17] which proposes the use of a system based on a deep learning model that utilises multi-layered hybrid architectures that automatically extract features inherent in traffic flow data, thus developing a module that extracts short-term temporal and spatial features based on the convolutional neural network (CNN) and the long short-term memory network (LSTM). They designed an attention mechanism capable of distinguishing and assigning weights to flow sequences at different times and proposed a bidirectional LSTM module (Bi-LSTM) that extracts periodic and weekly features in order to capture trends in traffic flow variation. The results presented from these experiments demonstrated a performance improvement in traffic prediction using a model combining Conv-LSTM and Bi-LSTM over other existing approaches.

The second study [18] proposes two new models, one that uses two long-short term memory (LSTM) units to extract temporal features of the traffic flow, followed by four dense layers that predict it, and another that uses two gated recurrent unit (GRU) units capable of extracting the temporal features followed by three layers for their prediction. Both models created can detect changes in trends and show good results in measuring performance, traffic and congestion in specific cases.

The third one [7] proposes an end-to-end deep learning architecture based on the combination of two models, a Conv-LSTM that extracts the spatial and temporal information of the traffic flow, and a bidirectional LSTM module to analyse the historical flow data to detect the periodicity of the traffic in order to detect possible patterns in the

data. This approach was applied on a real dataset and showed convincing results in improving prediction accuracy compared to existing studies.

According to the article [19] Bayesian Network Learning Framework for Travel Mode Identification Based on Cellular Signaling Data, identifying the mode of travel is essential for good traffic planning and management in order to optimise and relieve traffic congestion, proposing a Bayesian network to identify urban journeys via cell signalling, taking into account behavioural, personal and traffic environment attributes. An approach to learning the structure of the causal diagram of the Bayesian network that integrates information theory and probability theory, and one that is based on maximum a posteriori estimation, was proposed using geospatial data from the transport network, travel survey data and cell signalling data from the city of Kunshan. The article concluded by presenting a model to improve traffic planning and management with a very accurate traffic identification rate.

The study [20] presents an innovative methodology that enriches asynchronous time series data from a variety of sources to facilitate data enrichment and simulate traffic behaviour, using an XGBoost model, based on geographical coordinates and timestamp disparities, which makes it possible to improve the richness and granularity of the data from the different sources available. The study states that this methodology has simulation possibilities based on smart city datasets using asynchronous time series data integration and XGBoost for accurate traffic density prediction.

Work [21] states that building new roads to reduce growing congestion is not an effective way of doing this, and that alternatives need to be found to solve the congestion problem. One way presented was to disperse traffic demand by providing highly accurate traffic jam forecast information before drivers decide on their journey routes. A forecasting method was thus proposed that analyses population data in real time via the cellular network, using traffic demand from the population via linear regression and then travel time from predicted traffic demand via non-linear regression. The demonstrated results of this method were able to predict traffic in advance and quite accurately, and responses to a survey of drivers showed that there were changes in their behaviour when they saw the prediction results.

This study [22] uses GPS data from floating cars covering a large-scale region in Xi'an, China, in order to facilitate the processing of raw data to a more interpretable level with

practical applications, in the face of the accumulation of data in modern society that gives rise to proposals for increasingly data-based methods. A pre-processing method with specified time-frequency rules was proposed to reduce data dispersion and improve the quality of the original data, as well as non-parametric learning and survival analysis methods to assess and compare traffic congestion. Four different distributions (exponential, Weibull, log-normal and log-logistic) were also tested on randomly selected segments in order to adjust the AFT model and compare it to the Cox proportional hazards model, and the results showed the AFT (Lognormal) to be the most suitable for the chosen test section. The study concludes that the application of these models can potentially be effective in controlling and managing traffic in smart cities, showing a great improvement in the quality of the data and its credibility in assessing traffic conditions.

This study [23] tests the possibility of regional development of selected monocentric and polycentric urban areas in Germany in order to improve land use and transport planning and to test the risks of congestion in the transport network by applying data from cell phone networks and their mobility patterns. To do this, they explored data using visual and computational approaches, linking cell phone records to urban classifications and transport network data. The result of this approach provided a better understanding of the spatial and temporal patterns of mobility in different types of cities, and the potential of using cell phone network data to monitor spatial policies in land use and transport planning.

2.2. Search and inclusion tecniques

In this section, we will demonstrate the methodologies and strategies used to find relevant literature for our research. We used the Scopus database in order to better control the topics and population relevant to our study. The analysis was based on the title and abstract, taking into account the most relevant keywords for the topic. The search for these keywords was divided into three categories: Concept, Population and Context, and a time limit was set from 2015 to 2024. The concepts of "Data Analysis" or "Behavior Analysis" were thoroughly researched based on the populations "Smart cities", "Cellular network", "Roaming" or "City access", within the context of "Congestion".

This research, based on the limitations imposed, resulted in 83 documents when context, population and context are added together, as shown in table 1.

Table 1- Keywords definition

Concept	Population	Context	Limitations
Data Analysis Behavior Analysis	Smart Cities Cellular Network Roaming City access	Congestion	Only journal papers, articles, and reviews from 2015 to 2024
302,283 documents	153,135 documents	69,734 documents	
83 documents			

After analyzing how the keywords were defined, we studied the evolution by year of the studies identified, noting that they have become more popular in recent years, as illustrated in Figure 1.

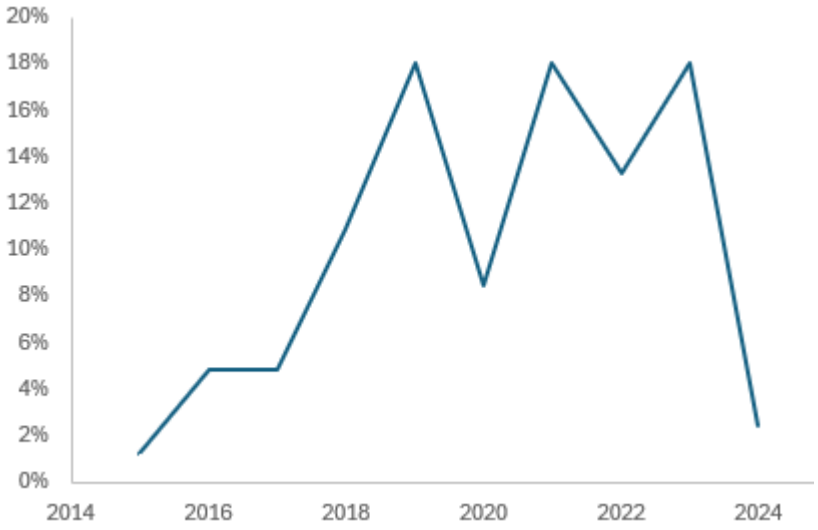

Figure 1- Evolution of relevant studies per year

CHAPTER 3

Knowledge Extraction Approach

3.1. Business Understanding

This phase aims to understand, collect, describe, investigate and validate the quality and soundness of the information and data. To ensure the proper execution of each of the activities, the data comprehension process was subdivided into three distinct phases:

a) Define business objectives
b) Data collection and evaluation
c) Compare and merge of the different databases

The dataset offered as part of this academic work was provided by the Lisbon City Council, which has launched a series of initiatives aimed at encouraging the academic world to develop analyses in the field of smart cities and the movement of people, and in this context has been taking steps to collect and provide information on the traffic flows that cause congestion.

3.2. Data description and understanding

The data used by customers of mobile phone operators is constantly being collected, obtaining various information and metrics that are used for different purposes, regardless of whether customers are using 2G, 3G, 4G or 5G internet.

Collecting this information creates a need for a base station in a certain location to cover a certain geographical area, interacting in real time with customers' mobile phones.

The work carried out in this dissertation was based on this interaction between the network and the users, receiving information on who enters and leaves the 11 access points in the city of Lisbon in the months between September 2021 and December 2022.

The data provided by Lisbon city Council (Câmara Municipal de Lisboa) via an agreement with a mobile operator is anonymous for legal and privacy reasons, making it impossible to analyze individual users. The data received is only aggregated in the grids representing the 11 access points to the city of Lisbon and is collected in 5-minute periods. In addition, we had access to weather data received from three different weather stations, which collect different information on different weather factors over a period of one hour.

To carry out this work, we were provided with two different types of databases:

- Commuter movements database: a set of individual monthly databases from September 2021 to December 2022, containing information of people entering and leaving the 11 main points of the city of Lisbon as folllow:
    - Eixo: The 11 access zones to the city;
    - Datetime: Year, month, day, hour, etc;
    - C12: No. of devices that entered the axes in 5 minutes;
    - C13: No. of devices that left the axes in 5 minutes.
- Meteorology database: that gives observations from IPMA meteorological stations:
    - Fecha: Information equivalent to Datetime;
    - Precacumulada: Accumulated precipitation given in ordered numerical values;
    - Humidity, radiation, temperature, etc...

Therefore, the following Table 2 and Table 3 provide a detailed description of each of the dimensions.

Table 2- Commuter movements dataset variables

ID	Variable Name	Variable Description	Variable Type
1	Eixo	Axis designation	Numeric
2	Datetime	Date of last reading	Date/Time
3	extract_year_2	Year of last reading	Numeric
4	extract_month_3	Month of last reading	Numeric
5	extract_day_4	Day of last reading	Numeric
6	C12	Number of devices that enter the identified axis during 5 minutes	Metric
7	C13	The number of devices that leave the identified axis for 5 minutes	Metric

Table 3- Meteorology dataset variables

ID	Variable Name	Variable Description	Variable Type
1	fecha	Date of last reading	Date/Time
2	estacion	Station no. - 01200535 - Lisboa Geofísico, 01200579- Lisboa Gago Coutinho, 01210762 - Lisboa Tapada da Ajuda	Text
3	humidade	Average relative humidity	Numeric, Percentage (%)
4	iddireccvento	Average wind direction - 0 to 9	Numeric, º - Degrees
5	intensidadeventokm	Average wind intensity	Numeric, Kilometre per hour (Km/h)
6	pressao	Atmospheric pressure	Numeric, Hectopascal (hPa)
7	radiacao	Solar radiation	Numeric, Watt per metre 2 - W/m2
8	temperatura	Temperature	Numeric, Degrees Celsius (oC)
9	precacumulada	Accumulated precipitation	Numeric, Millimetre (mm)
10	position	Coordinates and type of geographical entity (point, line or polygon)	Numeric

3.3. Data preparation

Data selection and cleaning: All duplicate and null rows were eliminated from the tables, representing around 2.86 of the data. First, we merged and cleaned the commuter movements individual monthly databases. We then added columns and variables that best matched the objectives set, that allow us to distinguish the data by public holidays, school calendars, rush hours and days of the week, which was essential in order to be able to identify the patterns of people entering and leaving the city of Lisbon. Next, in the meteorology table, the dtype of the Fecha column was changed to match the same as the Datetime column in the other table, so that we could merge the two. Also in the meteorology table, one of the important columns that was added was Nivel_Precipitacao, which divides the precipitation values in the precaccumulated column into four different levels for better observation, which indicate no precipitation, low precipitation, medium precipitation and high precipitation. Another important one was the Periodo_Letivo column, which divides the year into 5 school periods, where dates were set for each of the periods based on the actual school breaks and periods identified during the period under analysis.

Feature selection: After all the changes had been made, it was possible to carry out tests to demonstrate the model's performance, and thus better understand and predict traffic patterns, contributing to more effective and informed management of commuter movements on the city's main access routes. LSTM was used for these tests, as it was the one we found most effective for dealing with sequential and temporal data types such as ours. The positive results of using LSTMs in our data are shown in the following section.

Limitations: Given the nature of the data, it is only possible to analyse the data during the period from September 2021 to December 2022 and on the grids of the 11 identified access points to the city of Lisbon. With the data received, it is not possible to know the personal details of each individual travelling on the grids identified, nor to know their place of origin or destination. The large weight of the data used meant that the analyses and predictions made in the model applied took considerably longer. The time periods of data collection and the different types of data used between the two databases received made it difficult to aggregate the information from both databases.

3.4. Visualization

For a clearer visualisation of the data, PowerBi and Python were used to present the demonstrations that best fit the objectives initially set. The results and analysis of these visualisations are shown in Chapter 4.

CHAPTER 4

Visualizations

4.1. Insights and visualizations

In this chapter we will present the different analyses developed during the course of this work, where we will focus on answering the objectives initially outlined, identifying the times of day with the highest number of entries and exits, the busiest days of the week and months, and how school periods and the rainfall factor influence the variations in these commuter movements on the main access routes to the city of Lisbon.

4.2. What are the busiest times of day?

Figure 2 shows that the highest peak in the number of entries and exits is during the morning rush hour period, around 8:30, with a new peak appearing at 17:30, corresponding to the afternoon rush hour period reaching average values of 600 and 500 respectively. These values reach their lowest levels in the early hours of the morning, as low as 100, and after 21:00 around 200. The trend of incoming traffic is similar to that of outgoing traffic, both in terms of fluctuations and values. It can be seen that in the morning the number of entries is higher than the number of exits by around 50, while in the afternoon the opposite is true although the difference is much more subtle.

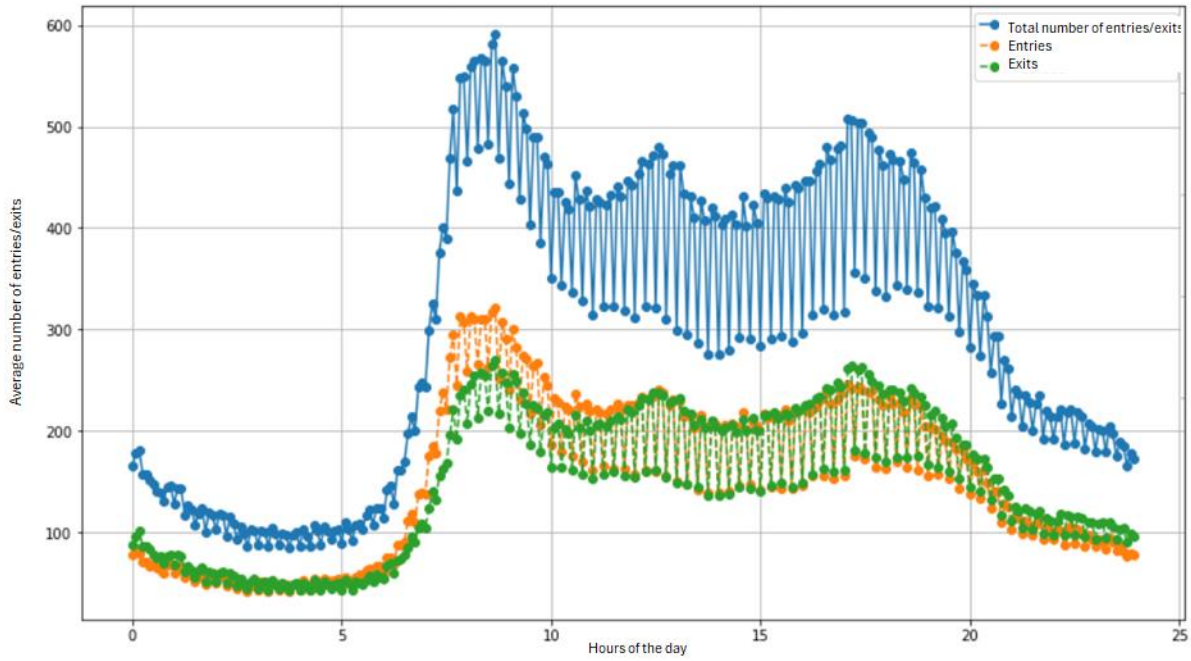

Figure 2- Average evolution of commuter movements throughout the day

4.3. Which days of the week have the most traffic?

Figure 3 shows the days of the week with the most traffic during the morning rush hour. It can be seen that Wednesday is the day with the highest average number of entries and exits, around 550, followed by the rest of the working days of the week with figures around 530. The big drop in traffic during the morning rush hour occurs at the weekend, with Sunday having the lowest average of around 200, and Saturday just over 300. On all days, the average number of entries is higher than the average number of exits, with a difference of around 50 during the week and a difference of less than 20 at the weekend.

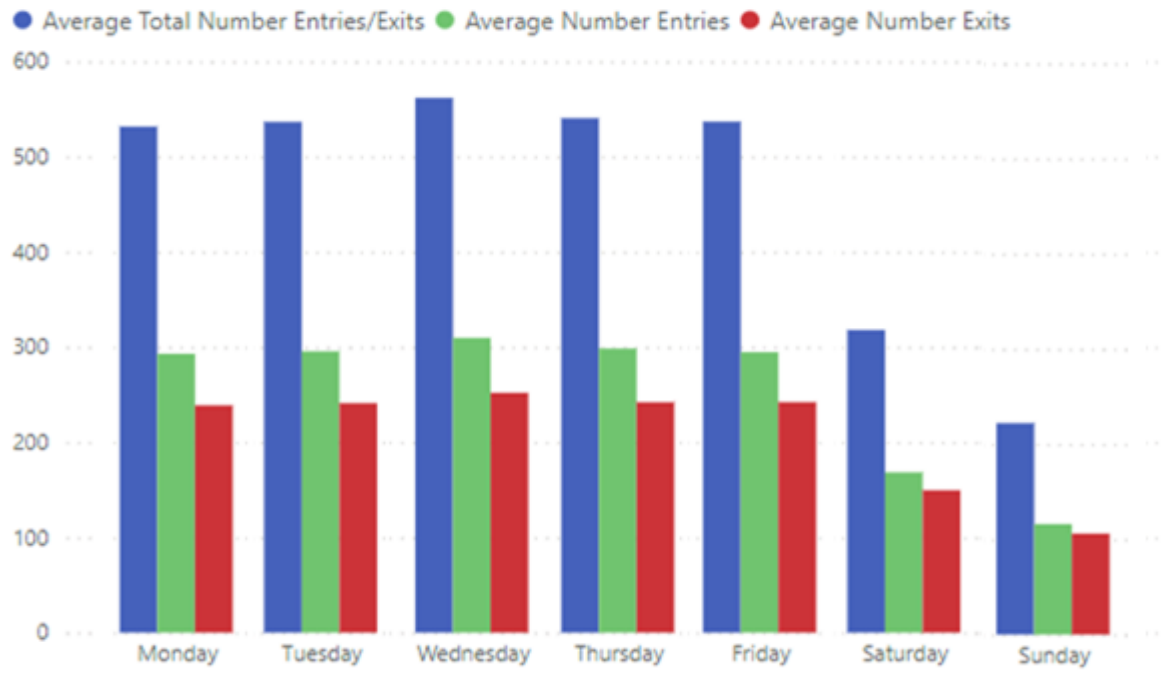

Figure 3- Average commuter movements by day of the week (morning rush hour)

Figure 4 shows the average number of entries and exits by day of the week, during the afternoon rush hour period. It can be seen that Friday is the day with the most traffic, with average values slightly above 400, followed by Sunday and Tuesday, which have values slightly below 400. Wednesday and Thursday are the days with the lowest average, with average values around 350. On most days, the number of departures is higher than the number of arrivals, with an average difference of no more than 30, with the exception of the weekend, when the number of arrivals was slightly higher.

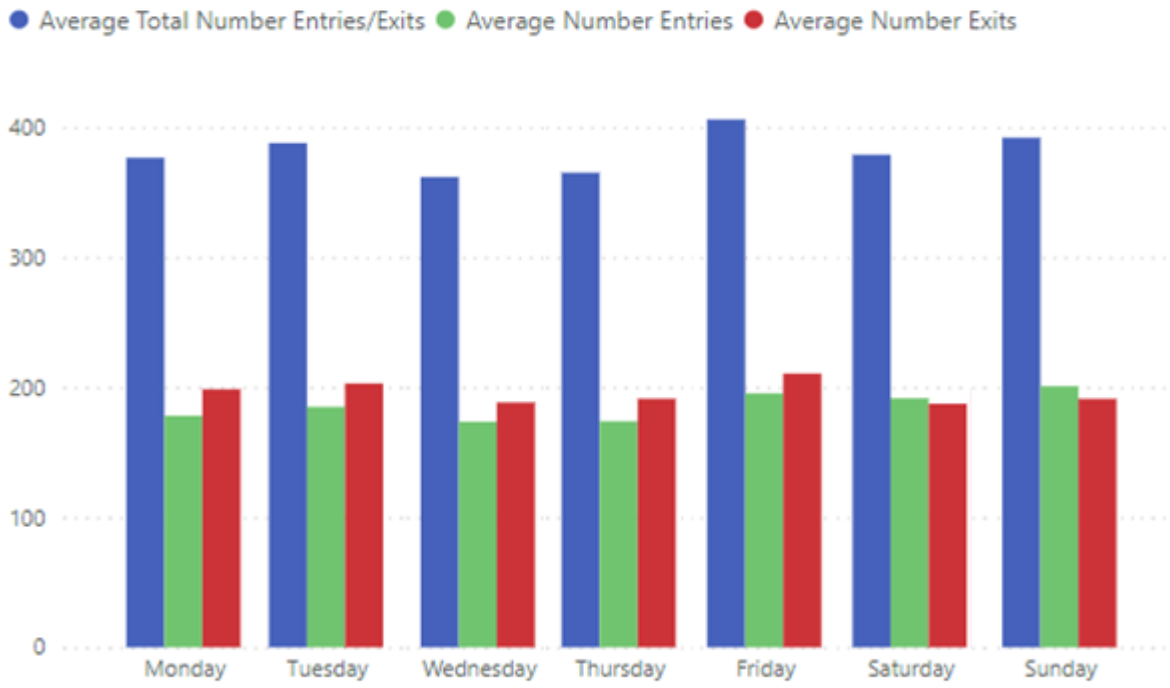

Figure 4- Average commuter movements by day of the week (afternoon rush hour)

Thus, the average number of entries and exits during the morning rush hour period is higher than in the afternoon on all working days of the week, except at the weekend, when there is a big rise in the average number. The weekend shows opposite trends when compared to both peak hours, with the lowest average value in the morning and the top 4 in the afternoon, with Sunday showing the highest average value in that period. Wednesday shows the same trend, with the highest value in the morning and the lowest in the afternoon, with a drop in the average value of around 200.

4.4. Which months have the most traffic?

Figure 5 shows the evolution of the average number of entries and exits by month during the morning rush hour period. It can be seen that the last four months of the year is the period with the highest average number of traffic, peaking in September at around 650, with an almost gradual decline in the remaining months until January. In February, there is a further increase, reaching an average value of around 450, but the following two months are the lowest of the year, as in August, with average values of no more than 250, with March recording the biggest drop of the year, at around 200. After these two months comes the second biggest rise of the year, with an average increase in the total number of arrivals and departures of around 150. The months of May to July show higher average figures than

January, with June showing the highest figure, at around 400. After the drop in August's average value, there was the biggest rise in the average number of traffic, around 400.

Throughout the year, the average number of entries was higher than the average number of exits, with a similar average difference in most months.

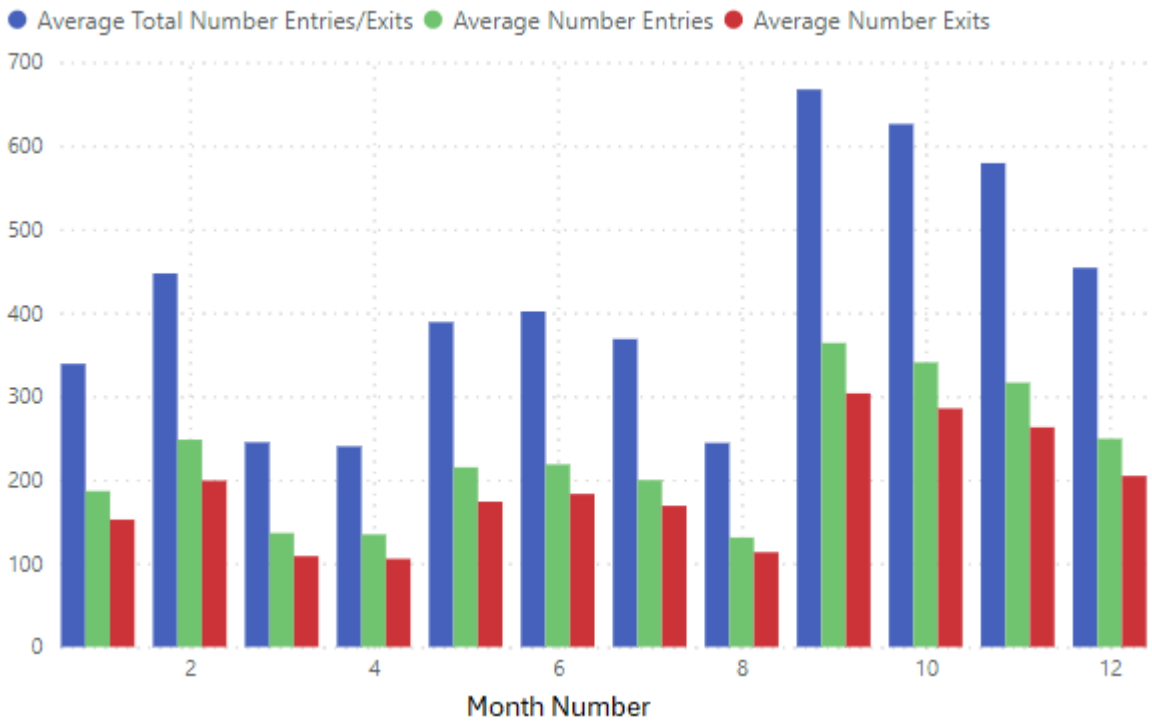

Figure 5- Average commuter movements by month (morning rush hour)

Figure 6 shows the evolution of the average number of entries and exits by month during the afternoon rush hour period. It can be seen that September and October are the months with the highest average traffic figures, at around 590 and 550 respectively. After September there was a drop in the average value until December, with November showing the biggest drop of the year, at around 170. The first two months of the year saw a further rise in the average value, with average values above 350, with a big drop in March. In the following months, there was a gradual rise until July, with an average value of around 300,

with a drop in August. After this month, there was a big increase in the average value, the highest of the year, above 300.

Throughout the year, the average number of exits was higher than the average number of entries, with a similar difference in average value in most months.

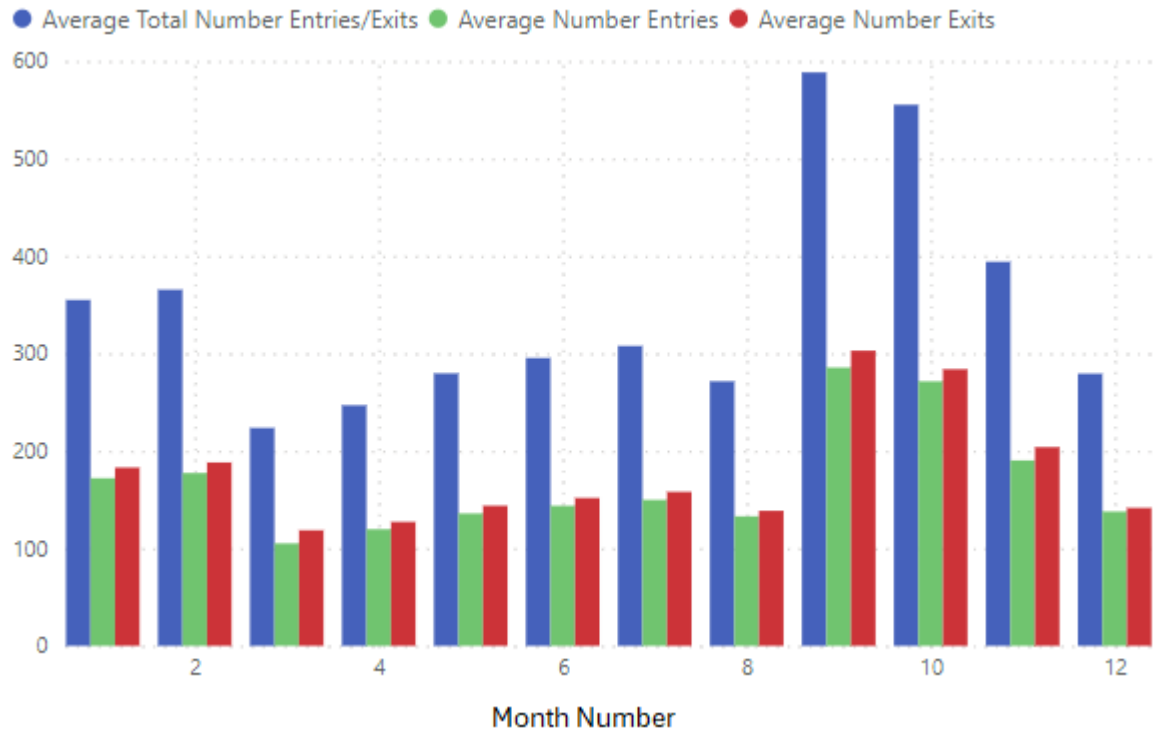

Figure 6- Average commuter movements by month (afternoon rush hour)

Thus, both peak hours show similar fluctuations when observed in relation to their average monthly traffic values, with both peaking in September, with a decrease in the remaining months of the year and with a new peak in February, and with a gradual increase during the first months of summer. The main differences between the peak hours are found in the average values presented for each month and in the sharp drops or rises in values between months.

4.5. What are the busiest school periods?

Figure 7 shows the average number of entries and exits during the morning rush hour period. It can be seen that the school term has the highest average total number of traffic, close to 500, followed by the winter holiday period and the exam period, which have average values close to 400. The summer and Easter holiday periods are the ones with the lowest average figures, around 260 and 230 respectively. As in previous analyses, in all

periods during the morning rush hour the number of arrivals is higher than the number of departures.

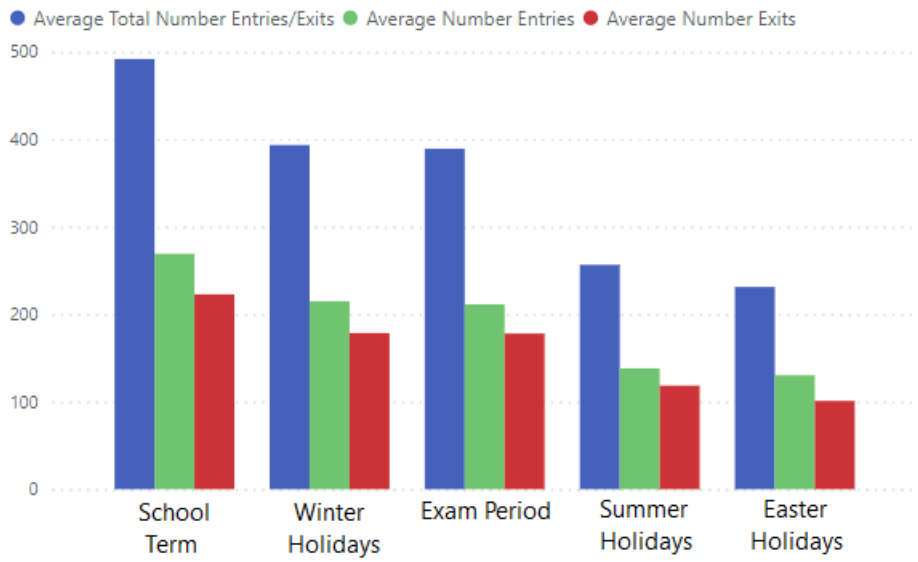

Figure 7- Average commuter movements by school term (morning rush hour)

Figure 8 shows the average number of entries and exits during the afternoon rush hour period. It can be seen that the school term has the highest total average number of traffic, with figures slightly above 400, followed by the exam period, with average figures around 300. Holiday periods have the lowest average traffic during the afternoon rush hour, with the winter and Easter holidays having the lowest average, around 240. In all the periods identified, the number of exits is slightly higher than the number of entries, and in some cases almost zero.

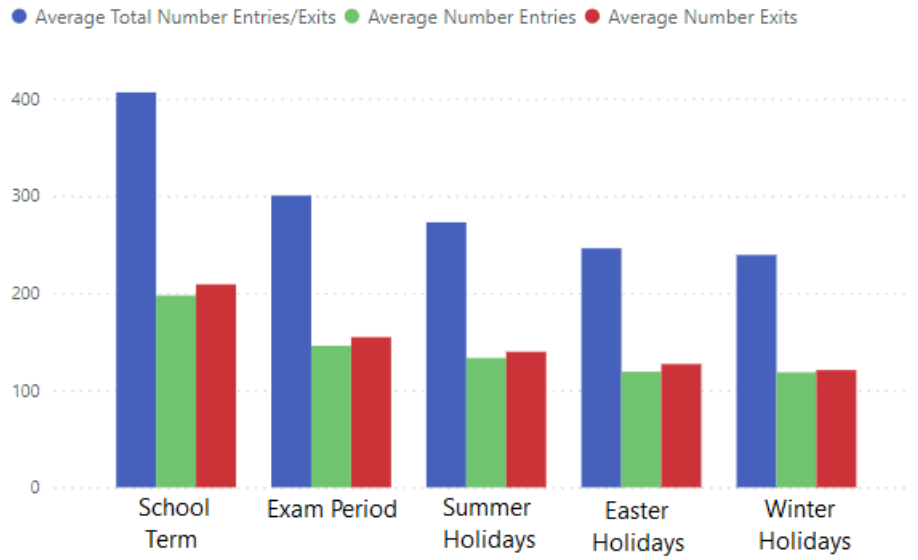

Figure 8- Average commuter movements by school term (afternoon rush hour)

4.6. How does rainfall influence traffic?

The Figure 9 shows the variation in average commuter movements by the different levels of rainfall as previously identified in chapter 3.3, during rush hour periods. It can be seen that the average number is highest when there is no precipitation of around 400 and its lowest value is when there is a low level of precipitation (level 1), around 420. then it rises sharply when level 2 precipitation occurs at around 400. At all levels the number of entrants is higher than the number of exits, with the biggest difference at level 2, at around 40.

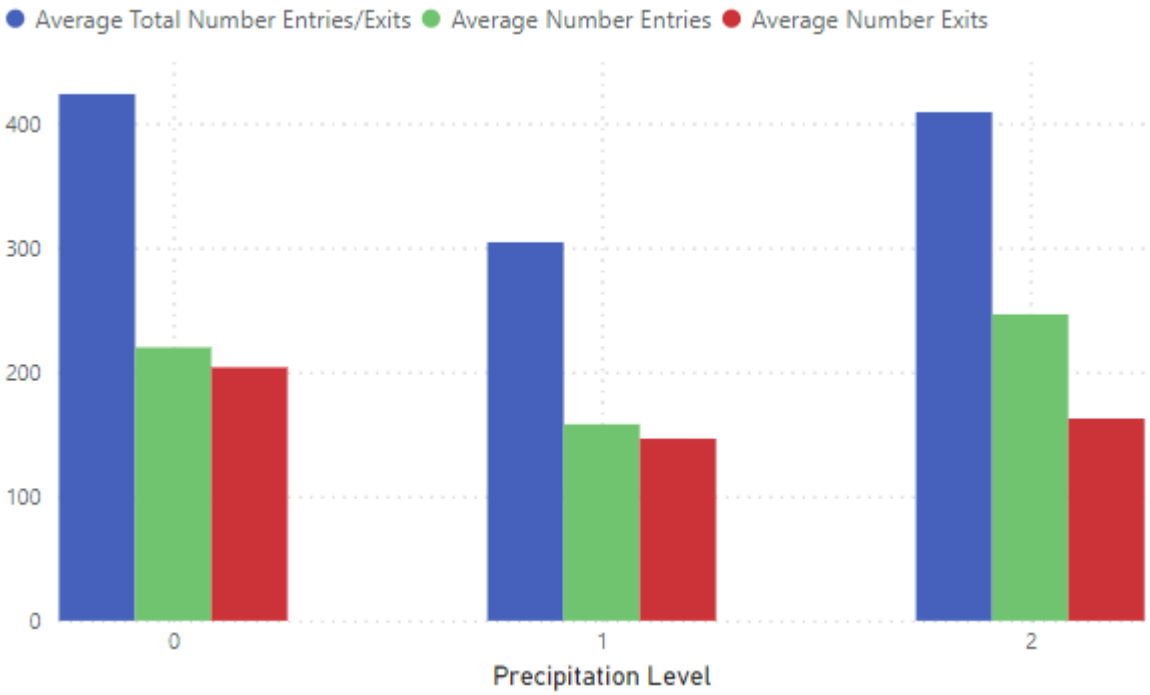

Figure 9- Average commuter movements by rainfall level

It is therefore possible to see different behaviours in the number of entries and exits by the existence or not of precipitation, however the precipitation levels initially defined are not as effective when only the peak hour periods are analysed, since of the four levels initially defined, only level 3 occurs during this period. When analysed in greater depth, it was found that during the afternoon rush hour period, precipitation level 2 did not occur. Therefore, the following two analyses will take into account the rainfall values initially given by the database.

Figure 10 shows that during the morning rush hour there was a decrease in the average number of entries and exits as the rainfall increased, with average traffic values of around 270 and 60 for rainfall values of 1 mm and 2 mm respectively. There were no occurrences of 3 mm of rainfall during the morning rush hour period, however, there was

an increase in the average number of traffic when the rainfall value of 4 mm was verified, with an average traffic value of slightly above 400. During all rainfall periods, the number of entries was higher than the number of exits, with the difference being almost zero when there was 2 mm of rainfall, and higher when there was 4 mm of rainfall, around 80.

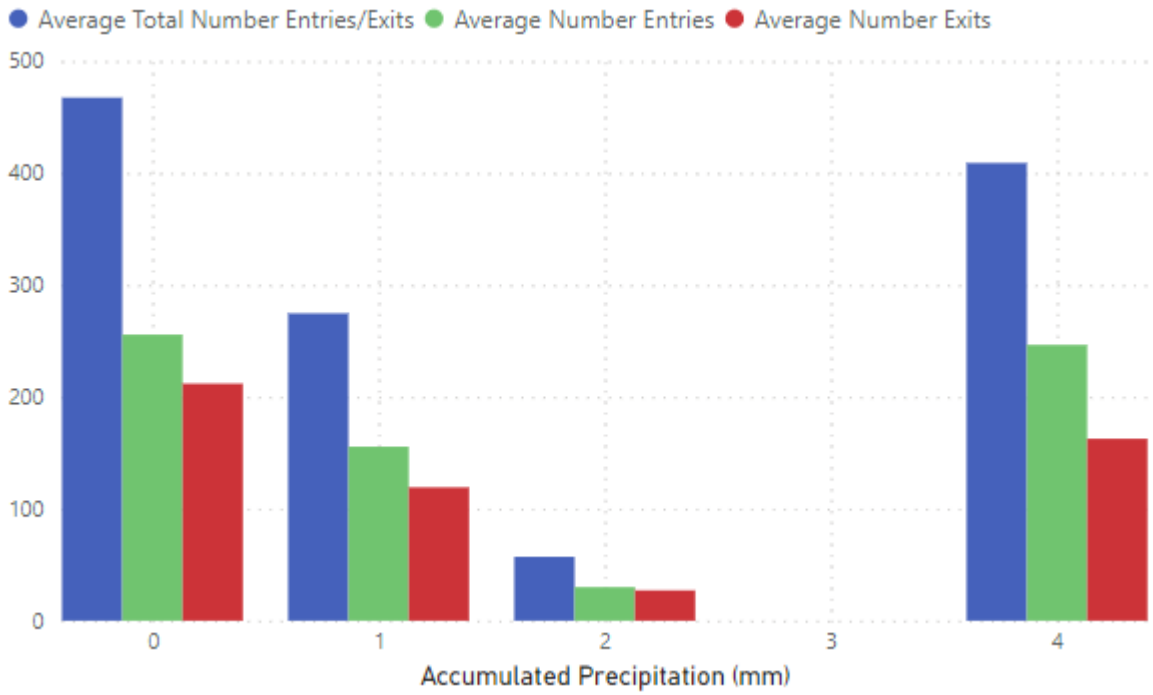

Figure 10- Average commuter movements by precacumulada values (morning rush hour period)

Figure 11 shows the average number of entries and exits for the different rainfall values during the afternoon rush hour period. The average number of entries and exits was highest when there was no rainfall, at around 380. There was a slight drop when there was 1 mm of rainfall, and a big drop when there was 2 mm of rainfall, with average values of around 360 and 250 respectively. There was a big rise in the average value for 3 mm of

precipitation, and although very close, it was still slightly below the average value of 380 presented when there was no precipitation.

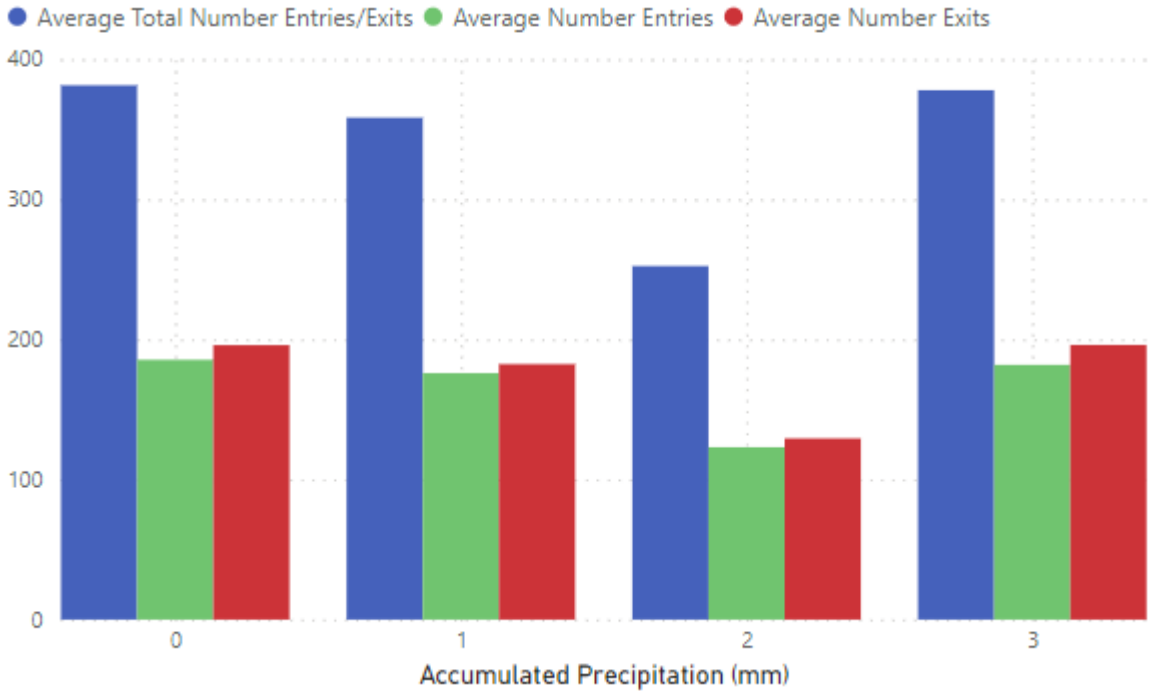

Figure 11- Average commuter movements by precacumulada values (afternoon rush hour period)

CHAPTER 5

Traffic Prediction and Detection

5.1. Predictions

In this chapter, we will look at the predictions made to the database to demonstrate their effectiveness. The analyses carried out were based on the different axes, for the different peak hours and considering the number of entries and exits. These forecast analyses were carried out in different phases, each with its own unique filters, and gave rise to graphical demonstrations of these forecasts and the loss values during the model's training and validation phase. These loss and val_loss values are important for understanding the performance of a model such as the LSTM used during its training phase, and the lower these values, the better the model's performance in making accurate forecasts. In each of the analyses carried out, 50 epochs were run to determine in which of the epochs the lowest loss values occurred. The graphs presented in this chapter, which are representative of the forecasts made, cover the period between September 2022 and December 2022, in order to check whether the model was able to predict the fluctuations in the number of entries and exits in the historical data of our study.

A first phase of analysis was carried out individually for each of the 11 entry shafts to the city of Lisbon, taking into account the total number of entries and exits, and for the rush hour period. The loss values obtained from these analyses were quite positive, showing loss values between 0.0001204 and 0.0147, at around 31 epochs (Table 4). Figure 12 graphically shows the results of the forecasts of the total number of entries and exits for each of the axes during both peak hour periods. It can be seen that in all the graphs the LTSM model was able to predict well the fluctuations in entries and exits during the period under analysis, detecting the rise in the average number of traffic around November, which represents the peak for most routes, and also detecting the almost universal drop for all routes around October.

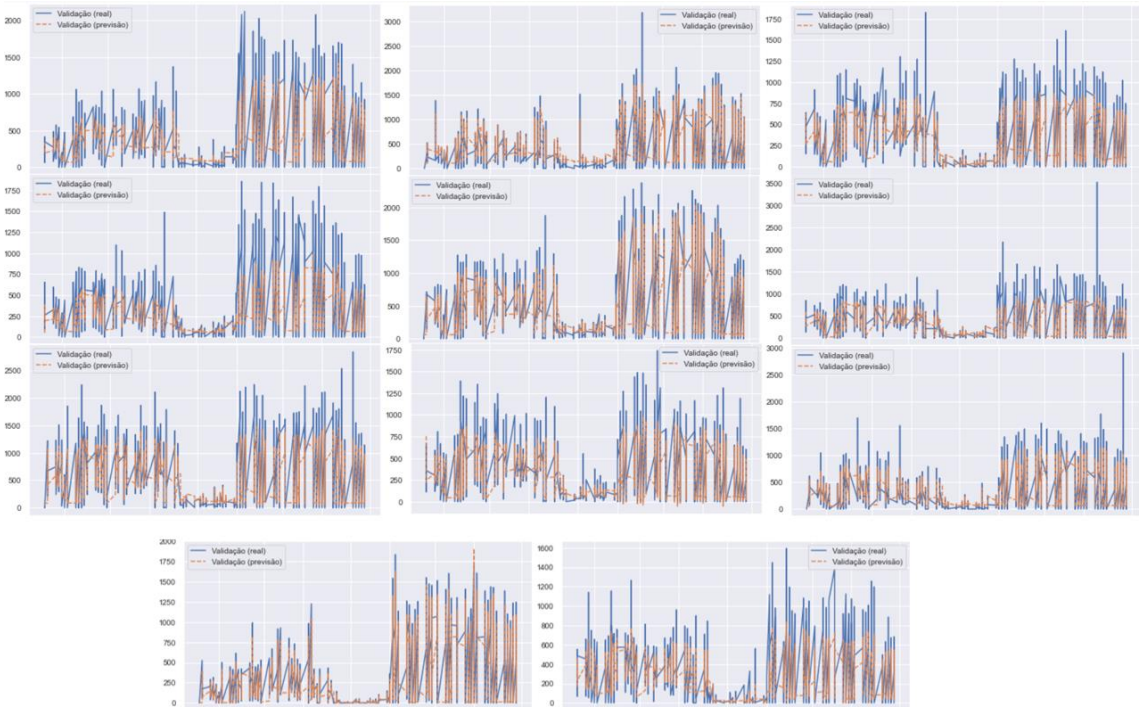

Figure 12- Trafic prediction in rush hour periods in the 11 Lisbon entry points

Table 4- Minimum loss and val_loss values shown for peak hour periods, for 50 epochs

	Rush Hour Periods		
Axes	Loss	Val_loss	Epoch
1	0,0029	0,0066	47
2	0,002	0,0024	22
3	0,00078196	0,003	30
4	0,0038	0,0147	43
5	0,0006489	0,0015	17
6	0,0015	0,0057	45
7	0,00063417	0,0013	19
8	0,0038	0,014	27
9	0,0013	0,002	26
10	0,0001332	0,0001204	25
11	0,0023	0,0142	41

In a second phase, forecasts were made for the same flow of entries and exits but taking into account only the morning rush hour period, in order to compare the results as can be seen in figure 13. In terms of performance, the model had positive loss and val_loss results, with values between 0.0015 and 0.0208 at around 31 epochs (Table 5). When analyzing the forecasts during the morning rush hour, it can be seen that there is greater differentiation in the fluctuations of traffic on each axis during the period under analysis.

However, the model proved capable of predicting the different fluctuations of the axes, successfully detecting these variations.

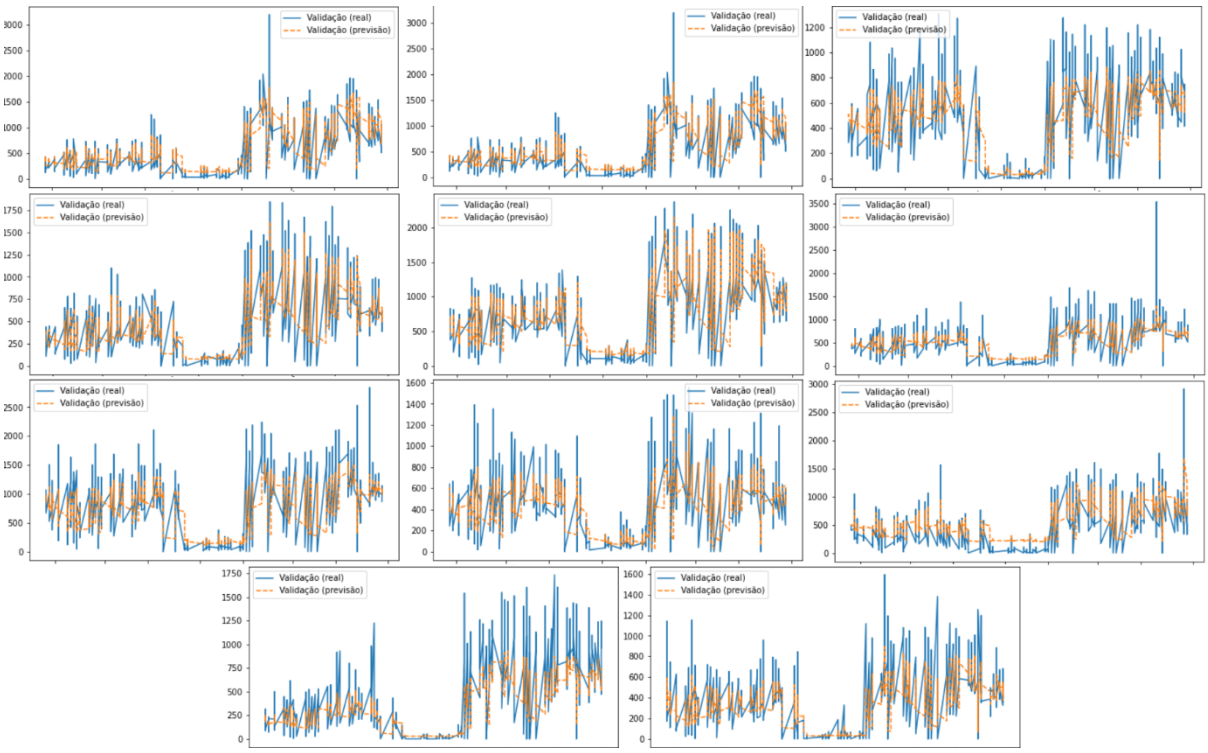

Figure 13- Trafic prediction in the morning rush hour periods in the 11 Lisbon entry points

Table 5- Minimum loss and val_loss values shown for the morning rush hour period, for 50 epochs

	Morning Rush Hour Period		
Axes	Loss	Val_loss	Epoch
1	0,0023	0,0022	33
2	0,0024	0,0023	24
3	0,004	0,0154	31
4	0,0044	0,0149	50
5	0,0005523	0,0015	44
6	0,0016	0,0056	46
7	0,00073817	0,0015	32
8	0,0065	0,0208	13
9	0,0012	0,0024	15
10	0,00070824	0,0029	29
11	0,0026	0,0174	23

New traffic flow forecasts were then made for the 11 axes but taking into account only the afternoon rush hour period, as shown in figure 14. In this analysis, the model showed the lowest loss and val_loss values, between 0.00064746 and 0.011 at around 36 epochs (Table 6), and was able to predict the new traffic variations. The axes generally

showed similar fluctuations and peaks in traffic, and the model was able to detect and represent them.

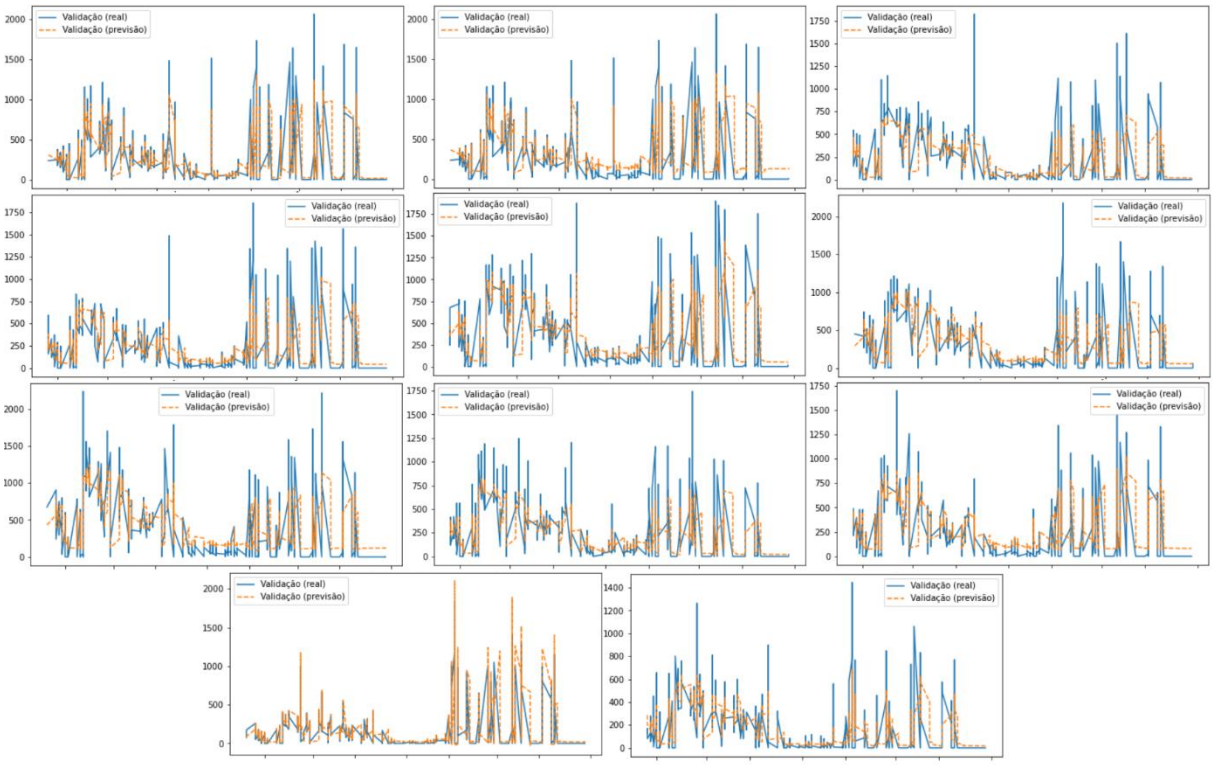

Figure 14- Trafic prediction in the afternoon rush hour periods in the 11 Lisbon entry points

Table 6- Minimum loss and val_loss values shown for the afternoon rush hour period, for 50 epochs

	Afternoon Rush Hour Period		
Axes	Loss	Val_loss	Epoch
1	0,0018	0,002	40
2	0,0018	0,0021	22
3	0,00072757	0,0023	27
4	0,003	0,011	48
5	0,0014	0,0031	32
6	0,002	0,0051	36
7	0,00066557	0,0009951	31
8	0,0029	0,0106	44
9	0,0043	0,0051	49
10	0,00021832	0,00064746	27
11	0,0023	0,0108	35

A model based on LSTMs was chosen because of its ability to make short-term forecasts using a large database. It was found in some of the studies analysed in chapter 2 that the LSTMs model proved superior at analysing and forecasting historical commuting data when compared to other types of models, which helped support the choice of LSTMs for our data.

The use of LSTMs proved to be very capable during the tests of this type of data, showing very low loss values in all the phases analysed, and successfully predicting the rises and falls in the number of inputs and outputs in the real values. We therefore agree that the LSTM model is ideal for short-term forecasting of historical traffic data.

CHAPTER 6

Conclusions and Future Work

6.1. Conclusions

Through the methods and techniques developed and used, we were able to respond to the objectives initially identified. We concluded that there are peaks in the average number of entries and exits during peak hour periods, with the morning peak hour showing higher values than the afternoon period, also showing that there is a higher average number of entries in the morning, while the opposite occurs in the afternoon, with a higher average number of exits.

During the morning rush hour, the weekdays have the highest average traffic numbers, with the highest value on Wednesdays and similar values on the other weekdays. The weekend has substantially lower average values than weekdays, reaching a minimum point on Sundays. There are big differences when compared to the afternoon rush hour period, with Wednesday becoming the day with the lowest average value, and Sunday the second highest, with Friday being the day with the highest average value. The difference between the highest and lowest values in the afternoon rush hour period is much lower than in the morning rush hour period.

There is a strong similarity in the way the average number of entries and exits evolves throughout the year in the morning and afternoon rush hour periods. Both show that the peak occurs during the month of September, with a gradual decline in the remaining months of the last quarter of the year. Likewise, in both peak hour periods, they spike again in February, with the following two months being the months with the lowest average during the year, and with another rise in the middle of the year, reaching another high around June/July, which marks the start of summer. During the morning rush hour, all the months have a higher average number of entries than exits, while the opposite is true during the afternoon rush hour.

In both morning and afternoon rush hour periods, the school period has the highest average values, but while the winter holidays have the second highest average during the morning rush hour, they have the lowest during the afternoon rush hour. As with previous analyses, during the morning rush hour, all periods show to have a higher average number of entries than exits, while the opposite is true during the afternoon rush hour.

We can conclude that the existence of precipitation influences the number of traffic observed, with the highest values when there is no precipitation, falling when precipitation is weak and rising again when precipitation levels reach the maximum value for each peak hour.

This study was carried out within the defined limitations, however the databases used allow a deeper study and analysis beyond our scope.

The mechanisms and methods of analysis used in the work are very helpful in traffic planning and to identify the periods of greatest congestion. The use of LSTMs has proven to be an effective type of neural network for predicting traffic flows at entry points into the city, capable of identifying patterns in the data.

The accuracy of the model applied allows us to have a lot of knowledge that goes beyond that studied in this work, which can be extended if users want to upload new updated data that follows the same structure as the data used. This allows decision-makers to have up-to-date and relevant information to analyse the data in order to plan and predict traffic, understanding the main causes of congestion, how it evolves over time and how it can be mitigated, improving the quality of life of individuals and the environment in the city of Lisbon.

6.2. Future work

For future work, and considering the topics covered and all the work carried out during this study, we present the following proposals or future enrichment:

a) Explore the fluctuations, time periods and factors that most influence congestion at access points to the city of Lisbon during off-peak periods.
b) Deepen the demonstrations by considering other factors that influence traffic and congestion other than those explored in the objectives of this work and on the basis of the data provided, by taking into account major cultural events that take place within the municipality of Lisbon, such as soccer matches or concerts that may take place inside or outside the rush hour period, which could lead to changes in forecasts and normal traffic flow.
c) Extend and deepen the analysis of the impact of climatic conditions on congestion beyond precipitation.

References

[1] A. Bhardwaj, S. R. Iyer, S. Ramesh, J. White, e L. Subramanian, «Understanding sudden traffic jams: From emergence to impact», *Dev. Eng.*, vol. 8, 2023, doi: 10.1016/j.deveng.2022.100105.

[2] M. Bawaneh e V. Simon, «Novel traffic congestion detection algorithms for smart city applications», *Concurr. Comput. Pract. Exp.*, vol. 35, n.º 5, 2023, doi: 10.1002/cpe.7563.

[3] A. Dutta *et al.*, «Intelligent Traffic Control System: Towards Smart City», apresentado na 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference, IEMCON 2019, 2019, pp. 1124–1129. doi: 10.1109/IEMCON.2019.8936188.

[4] Y.-H. Kuo, J. M. Y. Leung, e Y. Yan, «Public transport for smart cities: Recent innovations and future challenges», *Eur. J. Oper. Res.*, vol. 306, n.º 3, pp. 1001–1026, 2023, doi: 10.1016/j.ejor.2022.06.057.

[5] S. Godhbani, S. Elkosantini, W. Suh, e S. M. Lee, «ADL based Framework For Multimodal Data Fusion in Traffic Jam prediction», apresentado na International Conference on Software, Knowledge Information, Industrial Management and Applications, SKIMA, 2022, pp. 126–132. doi: 10.1109/SKIMA57145.2022.10029488.

[6] R. Goodspeed *et al.*, «Improving transit in small cities through collaborative and Data-driven scenario planning», *Case Stud. Transp. Policy*, vol. 11, 2023, doi: 10.1016/j.cstp.2023.100957.

[7] Y. Liu, H. Zheng, X. Feng, e Z. Chen, «Short-term traffic flow prediction with Conv-LSTM», apresentado na 2017 9th International Conference on Wireless Communications and Signal Processing, WCSP 2017 - Proceedings, 2017, pp. 1–6. doi: 10.1109/WCSP.2017.8171119.

[8] https://www.facebook.com/NoticiasAoMinuto, «Trabalhar ou estudar: Quantos entram e saem todos os dias de Lisboa?», Notícias ao Minuto. Acedido: 3 de julho de 2023. [Em linha]. Disponível em: https://www.noticiasaominuto.com/economia/2313884/trabalhar-ou-estudar-quantos-entram-e-saem-todos-os-dias-de-lisboa

[9] «PORDATA - Ambiente de Consulta». Acedido: 3 de julho de 2023. [Em linha]. Disponível em: https://www.pordata.pt/db/municipios/ambiente+de+consulta/tabela

[10] «Economia_de_Lisboa_em_Numeros_2022_PT.pdf».

[11] «Lisboa», MUNICÍPIO de LISBOA. Acedido: 1 de julho de 2023. [Em linha]. Disponível em: https://www.lisboa.pt/

[12] «DESAFIOS», LISBOA ABERTA. Acedido: 1 de julho de 2023. [Em linha]. Disponível em: https://lisboaaberta.cm-lisboa.pt/index.php/pt/lx-data-lab/desafios-teste

[13] «Instituto Português do Mar e da Atmosfera». Acedido: 1 de julho de 2023. [Em linha]. Disponível em: https://www.ipma.pt/pt/

[14] H. Zhu, K. Zhang, C. Wang, L. Jia, e S. Song, «The Impact of Road Functions on Road Congestions Based on POI Clustering: An Empirical Analysis in Xi'an, China», *J. Adv. Transp.*, vol. 2023, 2023, doi: 10.1155/2023/6144048.

[15] S. Zhou, C. Wei, C. Song, X. Pan, W. Chang, e L. Yang, «Short-Term Traffic Flow Prediction of the Smart City Using 5G Internet of Vehicles Based on Edge Computing», *IEEE Trans. Intell. Transp. Syst.*, vol. 24, n.º 2, pp. 2229–2238, 2023, doi: 10.1109/TITS.2022.3147845.

[16] Y. Lv, Y. Duan, W. Kang, Z. Li, e F.-Y. Wang, «Traffic Flow Prediction with Big Data: A Deep Learning Approach», *IEEE Trans. Intell. Transp. Syst.*, vol. 16, n.º 2, pp. 865–873, 2015, doi: 10.1109/TITS.2014.2345663.

[17] H. Zheng, F. Lin, X. Feng, e Y. Chen, «A Hybrid Deep Learning Model with Attention-Based Conv-LSTM Networks for Short-Term Traffic Flow Prediction», *IEEE Trans. Intell. Transp. Syst.*, vol. 22, n.º 11, pp. 6910–6920, 2021, doi: 10.1109/TITS.2020.2997352.

[18] M. S. Sawah, S. A. Taie, M. H. Ibrahim, e S. A. Hussein, «An accurate traffic flow prediction using long-short term memory and gated recurrent unit networks», *Bull. Electr. Eng. Inform.*, vol. 12, n.º 3, pp. 1806–1816, 2023, doi: 10.11591/eei.v12i3.5080.

[19] M. Wang, H. Liu, J. He, C. An, J. Xia, e Z. Lu, «Bayesian Network Learning Framework for Travel Mode Identification Based on Cellular Signaling Data», apresentado na IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, 2023, pp. 2991–2997. doi: 10.1109/ITSC57777.2023.10421870.

[20] E. Garcia, C. Serrat, e F. Xhafa, «Breaking Through the Traffic Congestion: Asynchronous Time Series Data Integration and Xgboost for Accurate Traffic Density Prediction», apresentado na Proceedings - Winter Simulation Conference, 2023, pp. 1747–1758. doi: 10.1109/WSC60868.2023.10408652.

[21] H. Akatsuka, Y. Kamata, T. Nagata, N. Komiya, M. Goto, e M. Terada, «Traffic Dispersion by Predicting Traffic Conditions based on Population Distribution», apresentado na Proceedings - 2021 IEEE International Conference on Big Data, Big Data 2021, 2021, pp. 1327–1336. doi: 10.1109/BigData52589.2021.9672061.
[22] W. Yuan, P. Wang, J. Yang, e Y. Meng, «An alternative reliability method to evaluate the regional traffic congestion from GPS data obtained from floating cars», *IET Smart Cities*, vol. 3, n.º 2, pp. 79–90, 2021, doi: 10.1049/smc2.12001.
[23] S. Fina, J. Joshi, e D. Wittowsky, «Monitoring travel patterns in German city regions with the help of mobile phone network data», *Int. J. Digit. Earth*, vol. 14, n.º 3, pp. 379–399, 2021, doi: 10.1080/17538947.2020.1836048.