How Tourists Move in a City

Catarina Castanheira
Bachelor in Data Science
Iscte University Institute of Lisbon
Lisboa, Portugal
catarina_cordeiro@iscte-iul.pt

Rita Almeida
Bachelor in Data Science
Iscte University Institute of Lisbon
Lisboa, Portugal
rita_alexandra_almeida@iscte-iul.pt

Duarte Marques
Bachelor in Data Science
Iscte University Institute of Lisbon
Lisboa, Portugal
duarte_marques@iscte-iul.pt

Guilherme Firmino
Bachelor in Data Science
Iscte University Institute of Lisbon
Lisboa, Portugal
guilherme_firmino@iscte-iul.pt

Luís B. Elvas
Instituto Universitário de Lisboa
(ISCTE-IUL), ISTAR-IUL, 1649-026
Lisboa, Portugal
0000-0002-7489-4380

Joao C Ferreira
Instituto Universitário de Lisboa
(ISCTE-IUL), ISTAR-IUL, 1649-026
Lisboa, Portugal
0000-0002-6662-0806

Abstract— **Little is known about the spatial behaviour of urban visitors, even though travellers create a massive amount of data (Big Data) when they visit cities. Using their behaviours, these data sources may be utilised to monitor their existence. Using Big Data, this article aims to analyse the digital footprint of urban visitors. Unlike others that rely on mobile device operators in Lisbon, this article establishes a partnership with Lisbon Municipality. We developed a Python approach to clean and prepare visualisation dashboards to understand tourists' movement in a city. The analysed case study (Lisbon) demonstrates how tourists tend to gather around a set of parishes during a specific time of the day during the months under study as well as how unusual circumstances, namely international events, impact their overall spatial behaviour.**

The aim of this paper is to identify these patterns so that local entities can better manage the allocation of resources in Lisbon.

Keywords—Roaming, tourists, Data Analytics, Cellular networks

I. Introduction

After a year marked by a severe pandemic scenario (2020) in which Portugal registered a sharp drop in international tourism, we witnessed a recovery in international tourism starting in the second half of 2021. [1] This trend continued in the first two months of 2022 (January and February) with a 769.2% increase in guests from abroad in February 2022 compared to February 2021 [1].

To get a better understanding of how tourists move in a city, we collaborated with the Lisbon City Council who provided us with data from a mobile carrier company – Vodafone, on the mobility of people (roaming and non-roaming) in the city of Lisbon based on cell phone data over September to December 2021 and January 2022. Later, we collected data of the weather from a national atmospheric organization – IPMA, in the same timeline and in the same city: Lisbon, Portugal.

Understanding human mobility and its patterns through cell phone data allows us to quickly and accurately analyse the mobility of those who move around the city of Lisbon and has great potential to support the organization of large events and the use of public transportation, for example.

Considering the theme previously addressed, the development of this study aims, using data provided by Lisbon City Hall and IPMA, to analyse the spatial-temporal behaviour of visitors to the city of Lisbon, more specifically people using cell phones in Roaming (possible tourists), to explain patterns observed by them:

- Understand main tourist places of interest, where tourists eat and where they stay;
- Identify main places of concentration and their variation over the months under study;
- Understand how mobility patterns in tourists change with the weather;
- Discover unusual mobility patterns in international events happening in Portugal during the months under study.

II. State of the Art

A. Search strategy and inclusion criteria

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) Methodology [2] was used to conduct a systematic literature review with the research question "What is the state of the art on the tourists' behaviour analysis and tourism mobility in smart cities?"

The database used for the search was Scopus, and the study took place between May 8 and 12, 2022; all of the findings had to be publications published within the past five years and written in English.

The search method was based on a single query with many research focuses.

This approach allowed researchers to count the number of articles that existed while considering the subject, context, and population under investigation.

Only papers were considered for this review.

Reviews, conference papers, workshops, books, editorials, and publications not linked to the topic were removed.

B. Study Selection

The title and abstract were used to make the first selection of articles, and in certain situations when that information was insufficient, the whole text was examined.

C. Data Extraction and Synthesis

Zotero and Microsoft Excel were used to handle and store the data. The information included the title, author, year, journal, topic area, keywords, and abstract.

A qualitative assessment was undertaken based on the results reported above for data synthesis and analysis.

Scopus was thoroughly searched for published work on the topic related with the concept "Data Analysis" or "Behaviour Analysis", the target population "Smart cities" or "Cellular network” or “Tourist” or “Roaming” and within a “Mobility” context of the study.

D. Results

The research was made by searching the existing literature regarding the concept, target population and the context of this study in Scopus detailed in Table 1. The query was made in the database and with the same restrictions and filters.

Table 1. Keywords Definition.

Concept	Population	Context	Limitations
Data Analysis Behavio#r Analysis	smart cities cellular network Touris* Roaming	Mobility	2018-2022
453.106 Documents	220.301 Documents	642.769 Documents	Only journal papers, articles, and reviews
3,156 Documents			
44 Documents			

We can see that when we use the keywords from each column in the query (Concept AND Population AND Context AND Limitations) resulting in 44 documents.
16 papers were retrieved after a manual procedure was completed in order to determine major subjects on their research questions and define the outcomes.
Year, region, RQ topic, and a brief description were all factors in our study systematization.

E. Study Characteristics

The 16 studies included in the review were chosen using the above-mentioned criteria.

The trend line in Figure 1 shows that the issue we're examining is growing in popularity, demonstrating its importance.

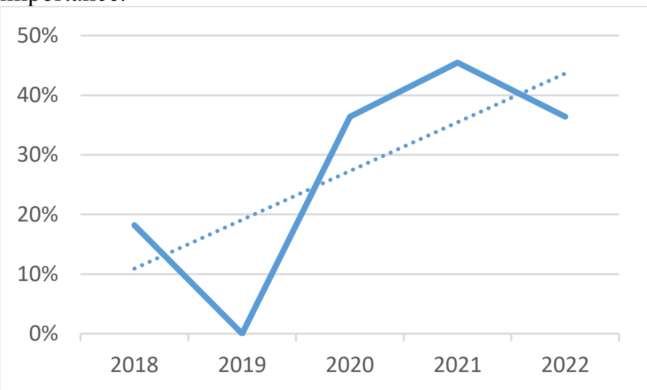

Fig 1. Evolution of the eligible studies by year

Given that the purpose of this study is to identify the use of tourist behaviour analysis and tourism mobility in Smart Cities, Table 2 and Figure 2 explain theoretically subjects mentioned in each of the evaluated papers, with a particular emphasis on the use of mobile phones and Behaviour Analysis on tourism and using mobile phones.

Figure 2 shows that most of the research focused on behaviours and used mobile phones and information and communications technology infrastructure (ICT).

Our research is predicated on both of these concepts since we not only investigate people's behaviour using the communication infrastructure of the city of Lisbon as an operator, but we also comprehend it and build a strategy to meet their demands.

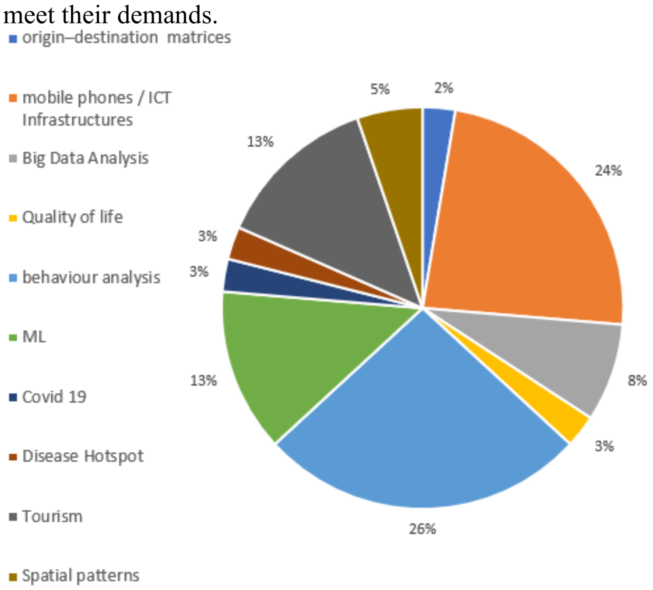

Table 2 summarises a more extensive examination of this review.

The issues were clearly described, and the writers of the publications did not need to be contacted for clarification.

The result categorization of the studies is not mutually exclusive, as they were assigned based on their existence or absence in the research.

Table 2. Keywords Definition.

Topic	Reference
(1) origin–destination matrices	[3]
(2) mobile phones / ICT Infrastructures	[3]–[11]
(3) Big Data Analysis	[3], [5], [10]
(4) Quality of life	[12]
(5) behaviour analysis	[4], [8], [10], [11], [13]–[18]
(6) ML	[4], [6], [7], [9], [13]
(7) Covid 19 / Disease Hotspot	[13]
(8) Tourism	[11], [13]–[15], [17]
(9) Spatial patterns	[14], [15]

After reading all of the publications, it was clear that the amount of behavioural research on tourist mobility has expanded dramatically in recent years all over the world.

Authors from [3] present a method for estimating origin–destination (O–D) matrices using passively obtained cellular network signalling data from millions of anonymous mobile phone users in the Rhône–Alpes region of France, enhancing and revolutionizing the field of travel demand and traffic flow modelling.

Still on the subject, the authors of study [5] can identify pedestrian hotspots and provide future traffic signal and street layout information to make the city more pedestrian friendly, as well as apply the knowledge gained to other data sets, such as bicycle traffic, to guide city infrastructure initiatives.

In a similar vein, but focusing on behaviour analysis, study [4] identifies a number of metrics for determining whether a person on the move is stationary, walking, or riding in a motorized private or public vehicle, with the goal of providing city users with personalized assistance messages for, among other things, sustainable mobility, health, and/or a better and more enjoyable life, with this applied to Tuscany and Florence. The goal of [8] in this chapter (combination of topics 2 and 5) is to study and compare the density of users in Shanghai city using Weibo geolocation data and univariate and bivariate density estimation approaches, such as point density and kernel density estimation (KDE), where the main findings are based on characteristics of users' spatial behaviour, such as the centre of activity based on check-ins, and the feasibility of using check-in data to explain the relationship between users and their social media accounts. Continuing in this vein, a research [10] based on long-term mobile phone data (from 2007 to 2012) of Beijing participants gives a means to visualize individual mobility patterns.

Study [6] aims to provide a taxonomy of 5G CN mobility prediction frameworks, from data gathering to model provisioning, while taking into account the 3GPP architecture and interfaces; and we provide two critical use cases in 5G CNs, in which the benefits of mobility predictions are assessed using information from real networks, whereas study [9] focuses on building a mobile sequential recommendation system to assist auto services (e.g., taxi drivers).

On the subject of behavioural analysis, study [16] presents an urban travel behaviour model and assesses its feasibility for creating a greener, cleaner environment for future generations, whereas study [18], based on a trip survey from the So Paulo Metropolitan Area, one of the world's busiest traffic locations, supplements a current bundling approach to enable multi-attribute trail datasets for the visual study of urban mobility, aiding in the identification and analysis of urban mobility.

In terms of quality of life, the authors of research [12] want to investigate the structural equation model of smart city elements that influence global management of world heritage sites as well as the quality of life for Thai tourists and inhabitants in Ayutthaya province.

Focusing on tourism, and behaviour analysis, author from [13] use machine learning to determine the most relevant parameters in influencing COVID-19 transmissions across different Chinese cities and clusters, researchers used a data-driven hierarchical modelling technique, being among this variables the “Number of tourists”. Following the same line, study [14] has the goal to assess the structure of tourist flows and \examine the variables that impact their regional distribution. Similarly, study [15] by using geographical and statistical analytic tools, the authors examine distinct intercity transportation patterns across different holidays and finds the driving factors, in order to optimize city hierarchical structure and allocate transportation resources. Study [7] using machine learning and the ICT, offer a position prediction system that takes into account both the spatial and temporal regularity of object movement. The object's historical trajectory data is utilized to derive personal trajectory patterns in order to determine possible future placements.

Using Airbnb data, author from [17] study visitors' mobility behaviour in relation to local public transportation access in tourist destinations. The author examines the attractiveness profile of 25 major tourist destination areas throughout the world using a "big data" analysis of the determinants of visitors' mobility behaviour and public transportation use in these tourist destinations.

Study [11] goal is to present novel techniques for studying pedestrian mobility aspects over the whole road network using the ICT and study the influence of visitor flows on the quality of life of locals and the preservation of cultural assets is exemplified by Venice.

III. METHODOLOGY: CRISP-DM

Cross Industry Standard Process for Data Mining (CRISP-DM), the methodological approach for the development of our project, is based on a standard method for designing data mining projects to reduce costs, increase reliability, execution, and manageability, making the data mining process more efficient [19].

However, for this project, given the data in question and our main goals, we opted for a modified version of this methodology consisting of 4 phases (see Fig 2):

- Data Understanding
- Data Preparation
- Analysis
- Visualisation

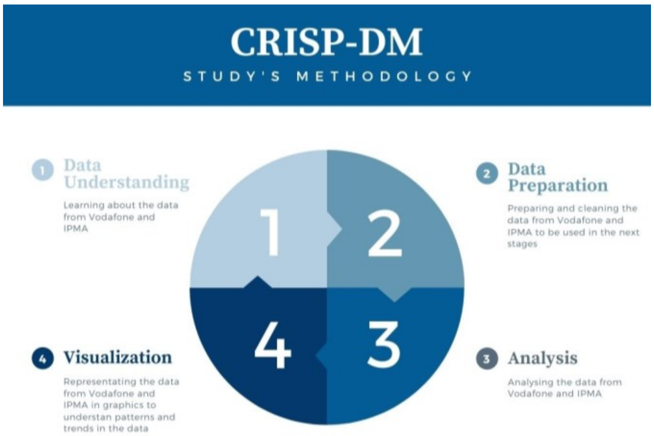

Fig 2 - Study’s Methodology

A. Data Understanding

After we gathered the data for our study, we explored it thoroughly and investigated each variable to understand the potential of our data and how we could maximize the added value of this study.

As stated previously, our main goal was to understand mobility patterns in tourists. To achieve that, we were given, by the Lisbon City Council, a dataset related to the mobility of people (roaming and non-roaming) in the city of Lisbon based on cell phone data, created by Vodafone, a British multinational telecommunications company. As it was explained to us, the data was collected from people who have Vodafone as a mobile carrier and later anonymized. To get a more accurate representation of the mobility of the all the people who circulated in Lisbon between September 2021 and January 2022, Vodafone extrapolated the data, originating the dataset we have available now (Vodafone dataset).

To get more accurate insights on the mobility of international tourists and how weather conditions change their mobility patterns, we reached out to the Portuguese Sea and Atmosphere Institute (IPMA) and requested data on the meteorological conditions in Lisbon between the months of September 2021 and February 2022.

There are two different datasets (Vodafone dataset and IPMA dataset), so they will be addressed separately.

1) Vodafone Dataset

The dataset provided by Vodafone was divided into several files in csv format separated by months, having 26 variables and 126 443 863 records in total, see Table I. As for the size of observations per month, see Table II.

TABLE I. VODAFONE DATASET VARIABLES

ID	Variable Name	Variable Description	Variable Type
0	Grid_ID	Number of grids There are 3743 squares of 200 by 200 meters in order to cover the metropolitan area of Lisbon	Nominal
1	Datetime	Time and date of occurrence	Datetime
2	C1	Number of distinct terminals counted on each grid cell during the 5 minute period – Measured every 5 minutes	Metric
3	C2	Number of distinct terminals in roaming counted on each grid cell during the 5 minute period– Measured every 5 minutes	Metric
4	C3	No. of distinct terminals that remained in the grid cell counted at the end of each 5 minute period	Metric
5	C4	No. of distinct terminals in roaming that remained in the grid cell counted at the end of each 5 minute period	Metric
6	C5	No. of distinct terminals entering the grid	Metric
7	C6	Terminals leaving the grid – These are the distinct terminals that left the grid. The calculation is made using the previous 5-minute interval as reference, also considering the crossings of the grid in the same interval	Metric
8	C7	Number of entries of distinct terminals, in roaming, in the grid	Metric
9	C8	Number of exits of distinct terminals, in roaming, in the grid	Metric
10	C9	Total no. of distinct terminals with active data connection in the grid cell – Measurement every 5 minutes	Metric
11	C10	Total no. of distinct terminals, in roaming, with active data connection in the grid cell – Measurement every 5 minutes	Metric
12	C11	No. of voices calls originating from the grid cell	Metric
13	C12	Entering the city: No. of devices that for 5 minutes enter the 11 street sections considered for analysis. For this purpose, a section of track is considered to be a route with	Metric

TABLE II. NUMBER OF OBSERVATION PER MONTH (IPMA DATASET)

MONTH	OBSERVATION NUMBERS
September 2021	17 233 318
October 2021	32 627 337
November 2021	21 619 292
December 2021	33 121 658
January 2021	33 344 624

2) IPMA Dataset

We performed an aggregation of the IPMA dataset by months and obtained a database consisting of 19 variables and 392 290 observations. The variables in this database are presented in Table II, as well as their descriptions:

TABLE III. IPMA DATASET VARIABLES

ID	Variable Name	Variable Description	Variable Type
0	Date_time	Date and time	Datetime
1	Entity_id	Entity type and id of the weather station (contains the values of the entity_type and station variables)	Nominal
2	Entity_location	Location of the station	Coordinates
3	Entity_ts	-	-
4	Entity_type	Entity type (present in variable entity_id)	Nominal
5	Station	Station id (present in variable entity_id)	Nominal
6	Fecha	Date and time of occurrence	Datetime
7	Fiware_service	firmware	-
8	Fiware_servicepat h	Path of firmware service	-
9	Humidity	Relative humidity	Numeric
10	Wind_direction	Wind direction (represented by heading classes in weather reports: N, NE, E, SE, etc.)	Ordinal
11	WindIntensity_txt	Wind intensity (coding of the values of the variable Wind_direction)	Ordinal
12	WindIntensityKm	Wind intensity in km/h	Metric
13	Position	Location of station	Coordinates
14	PrecAccumulated	Accumulated precipitation in millimetres	Metric
15	Pressure	Atmospheric pressure in hPa	Metric
16	Radiation	Solar radiation	Metric
17	Temperature	Temperature in Celsius	Metric
18	Validity_ts	-	-

B. *Data Preparation*

Now that we have a better understanding of the data, we move to the second phase. This phase consists of four subphases: data selection, data cleaning, resource selection and data integration. Originally the dataset was spread over several files in csv format, each month consisting of 4 to 9 files of the same format. To handle the data more efficiently on personal computers and to proceed with preparation of the data, we decided to merge the csv files by months, rather than compiling all the files provided into one file. Later, we did the same to the IPMA dataset which was divided by months.

1) *Vodafone Dataset*

a) *Data Selection*

Mobile roaming data is obtained by means of radio waves, which are sent and received by the telecommunications base station and automatically stored in the memory or in the log files of the mobile network operators, in this case Vodafone. When a cell phone is registered in one country but used in another, its user can be recognized as a potential tourist, and the corresponding information such as the country of origin and location coordinates are registered as mobile roaming.

The information collected by Vodafone was aggregated over 3743 squares of 200 by 200 meters, with no values of less than 10 devices reported, and collected in 5-minute periods. The data becomes available after a processing period of approximately 45 minutes. This information is very important to study the mobility of tourists and is one of the most accurate.

b) *Data Cleaning*

In the primary clean-up of the dataset, we discarded missing values and removed duplicate rows. When that was complete, we moved on to the next subphase.

c) *Resource Selection*

One of the first things we did after cleaning the dataset was selecting the variables weren't of interest to our objective and eliminating them from the dataset. Thus, the original dataset was minimized to only a few variables of interest for our objectives.

Subsequently we created a dataset with only the mobility data of people in roaming (tourists) from the Vodafone dataset. We achieved that by keeping only the variables related to people in roaming.

d) *Data Integration*

The variable Datetime was in the format %Y-%m-%dT%H:%M:%S.%fZ. In order to facilitate its visualisation and also the manipulation of the data, we converted this variable from the object format to datetime, through the datetime library, and subsequently created separate columns for the date, time and created a new column with the corresponding day of the week.

As the goal of our work focuses on the mobility of people, it was also important to distinguish between holidays and weekdays/weekends in our dataset. To do this we used the holidays library, and only marked holidays on weekdays, since at the weekend they will not have much impact on mobility in general. Thus, using Python's Numpy library, a matrix was created for both the Weekday column - 1 is weekday, 0 weekend, and the new Holidays column - 1 is a holiday, 0 is not a holiday.

To check some events during the different times of the day, a column with distinct time intervals was also created.

To the Vodafone dataset we coupled the Vodafone Grid dataset, to have information about the parish and latitudes and longitudes of each Grid_ID. From the junction of this dataset, it was possible to build new columns to facilitate a posterior data analysis and visualisation in PowerBI. A column with zones of Lisbon was then created, in which the 24 parishes of Lisbon were grouped into 5 distinct zones (visible in fig 2), according to the administrative reorganization of the parishes in 2012, namely (https://www.am-lisboa.pt/451600/1/008910,000505/index.html):

- North Zone (Green Zone) - Santa Clara, Lumiar, Carnide, São Domingos de Benfica, Benfica;
- Western Zone (Yellow Zone) - Alcântara, Ajuda, Belém;
- Center Zone (Orange Zone) - Campolide, Alvalade, Avenidas Novas, Santo António, Arroios, Areeiro;
- Historic Downtown Area (Purple Zone)- Campo de Ourique, Estrela, Misericórdia, Santa Maria Maior, São Vicente, Penha de França;
- Eastern Zone (Blue Zone) - Beato, Marvila, Olivais, Parque das Nações.

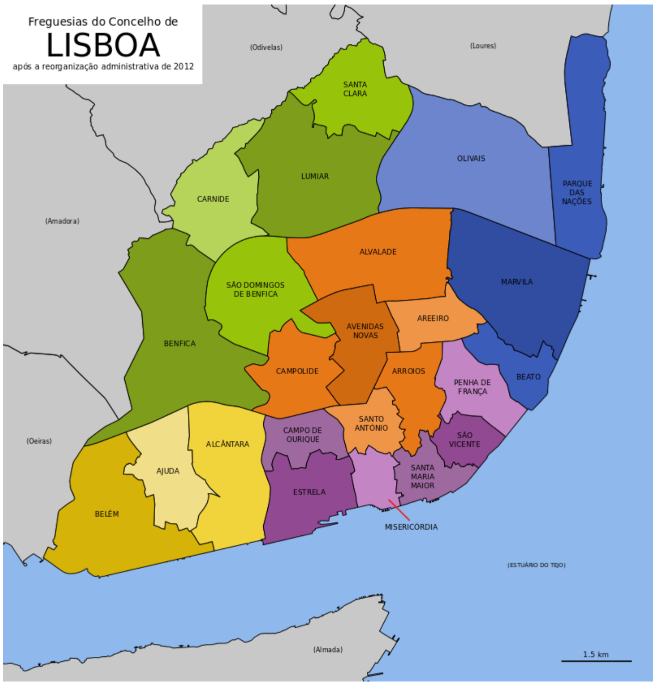

Fig 3 - Municipalities of the County of Lisbon

Regarding a variable (D1), which presents the TOP10 of equipment origin (cell phones) by order of representativeness, a split was applied in order to separate the countries that belonged to that column by distinct columns: TOP1 to TOP10, and then decreasing the TOP10 to only one TOP3 of representativeness of the countries in that location, so that instead of creating 10 additional columns, only 3 columns were created, making data manipulation more efficient.

For a facilitation of the visualisation of the mobility of tourists through the city of Lisbon, the countries represented in the columns TOP1 to TOP3 were grouped by territorial zones, except the main nationalities of tourists in Portugal: Spain, France, United Kingdom (Ref[a] Tourism Statistics

2020) and other countries that in our perspective it would not make sense to be grouped together. Three new columns were then created (zonasnacionalidade1, zonasnacionalidade2 and zonasnacionalidade3), which represent the TOP1, TOP2 and TOP3, respectively, of different nationality groupings, as follows:

TABLE IV. COUNTRY GROUPS

Country	Group
Spain	Spain
France	France
Germany	Germany
Italy	Italy
U. S. A	U. S. A
South Africa	South Africa
Russia, Belarus	Eastern Europe
Bulgaria, Romania, Lithuania, Estonia, Latvia, Ukraine, Slovakia	Eastern Europe
Denmark, Sweden, Finland, Norway, Iceland, Faroe Islands	Nordic Europe
Belgium, Luxembourg, Andorra	Central Europe
Malta, Slovenia, Croatia, Greece, Serbia, Bosnia and Herzegovina, Albania, Montenegro, Macedonia.	Southern Europe
Hungary, Poland, Czech Republic	Visegrad Group
Austria, Switzerland, Liechtenstein	Alpine Countries
Kuwait, Iraq, Afghanistan, Israel, United Arab Emirates, Turkey, Saudi Arabia, Cyprus, Iran, Egypt, Qatar, Oman, Jordan, Bahrain	Middle East
Singapore, Indonesia, Georgia, Philippines, Malaysia, Thailand, Armenia, Vietnam	Southwest Asia
Nigeria, Azerbaijan, Ghana, Cameroon, Ivory Coast, Benin	West Africa
Tunisia, Morocco, Kenya, Gibraltar, Mozambique, Uganda, Algeria, Mauritius, Reunion	North and East Africa
India, Bangladesh, Pakistan, Kazakhstan, Sri Lanka	South, Central and South Asia
Hong Kong, Japan, South Korea, Macau, Taiwan	Eastern Asia
Costa Rica, Panama, Puerto Rico, Mexico, Guatemala, Dominican Republic, Bermuda, Nicaragua	Central and North America
Brazil, Peru, Uruguay, Chile, Paraguay, Netherlands Antilles, French Guyana	South America
Australia, New Zealand	New Zealand and Australia
Ireland, United Kingdom	Ireland and United Kingdom
Nan, None	Unknown

2) *IPMA Dataset*

a) *Data Selection*

IPMA, the Portuguese Sea and Atmosphere Institute holds the largest national meteorological observation infrastructure network which generates data about the weather conditions in real-time. This data is then grouped into datasets to be studied and analysed by professionals qualified to know the weather and its variations, to aid in prediction models, time series, i.e. In this case the data is being used to complement the main dataset due to its importance in understanding the mobility patterns in tourists.

b) *Data Cleaning*

To clean the dataset, we discarded missing values, removed duplicate rows, and deleted all rows which had an observation with the number -99 (Error Code). After that, we moved on to the next subphase.

c) *Resource Selection*

We proceeded to eliminate the variables that weren't of interest to our objective and the variables that had no variance. Thus, the original dataset was minimized to only a few variables of interest (see table V), the most important ones being humidity, wind intensity, accumulated precipitation, and temperature.

TABLE V. REMAINING VARIABLES (IPMA DATASET)

ID	Variable Name	Variable Description	Variable Type
0	Day	Day of observation	Datetime
1	Hour	Hour of observation	Numeric
2	Month	Month of observation	Numeric
3	Month_txt	Month of observation in text format	Nominal
4	Station	Station where the data was collected from	Nominal
5	Position	Coordinates of the Station	Coordinates
6	Humidity	Relative Humidity in %	Numeric
7	Humidity_txt	Categorized version of the variable Humidity	Categorical
8	WindIntensity	Wind intensity Index used by IPMA	Ordinal
9	WindIntensity_txt	Categorized version of the variable WindIntensity	Categorical
10	WindIntensityKm	Wind Intensity in km/h	Metric
11	PrecAccumulated	Accumulated Precipitation in mm	Metric
12	PrecAccumulated_txt	Categorized version of the variable PrecAccumulated	Categorical
13	Temperature	Temperature in °C	Metric
14	Temperature_txt	Categorized version of the variable Temperature	Categorical

d) *Data Integration*

The variable fecha was in the format %Y-%m-%dT%H:%M:%S.%fZ. For the same reason as the Vodafone dataset, we converted this variable from the object format to datetime, through the datetime library, and subsequently created separate columns for the month, hour, and day of the observation. Having completed that we then created new columns that were the categorical version of our numeric variables of interest (see Table VI), basing our categories in the metadata from IPMA and the measuring system of each variable, and created a text version of the variable month to aid in the visualisation.

TABLE VI. NEW CATEGORICAL VARIABLES (IPMA DATASET)

Variable Name	Categories	Scale
Humidity_txt	No humidity	0–5 (%)
	Low humidity	6–20 (%)
	Moderate humidity	21–60 (%)
	High humidity	61–100 (%)
WindIntensity_txt (Index used by IPMA from 0 to 12)	No wind	0-1
	Low wind	1-4
	Moderate Wind	4-7
	Intense wind	7-12
AccPrecipitation_txt	No rain	0-5 (%)
	Low rain	6-20 (%)
	Moderate rain	21-60 (%)
	Abundant rain	61-100 (%)
Temperature_txt	Very Cold	-5-0 (°C)
	Cold	1-20 (°C)
	Pleasant	21-28 (°C)
	Warm	29-40 (°C)

IV. VISUALISATION

After the data analysis, we moved on to data visualisation. In this stage, we represented the data in graphics, as it's a quick way to see trends and patterns in the mobility of the tourists and to focus on the most important points. Graphs and charts let us explore and learn more about data [20].

We found that of the months under review, October had the most tourists (see Fig 4).

Fig 4 - Tourists in Lisbon

As for the presence of tourists in the various periods of the day, the month of October was analysed since it is the month that presents more tourist visits, as seen in the previous graph (see Fig 4). We found that in the lunch hour period the parishes that tourists frequent the most are Belém, Alvalade and Parque das Nações; in the afternoon and evening periods the most visited parishes are Belém, Alvalade and Avenidas Novas; finally, during the early morning hours the most frequented parishes are Avenidas Novas, Belém and Arroios.

Good data visualisations also makes it easier to communicate our ideas and insights to other people [20]. With PowerBI we represented each parish and used the data available to visualise the movement of tourists and to identify major patterns.

A) Case 1 – Mobility Patterns in Different Weather

One interesting case is the influence of tourists' movements based on weather conditions and the places they prefer to go to when it's raining and when it's sunny. In Fig 4 and 5 we represented tourists' movements on sunny and rainy weeks to depict their movements. When it is raining tourists tend to concentrate in certain places (The historic downtown area, the airport, near Carnide where we found a lot of hostels and apartments for rent and in the Eastern zone). Tourists tend to stick to monuments, their places of stay and shopping centres in rainy days (see Fig 5). When it's sunny, rather than concentrating in some places, tourists tend to spread out more and visit more areas than the "hotspots" (see Fig 6). It's clearly visible the effects of the weather on the mobility of tourists.

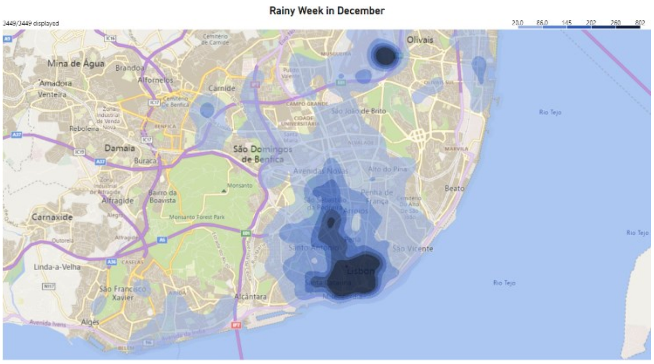

Fig 5 – Tourists' mobility on a rainy week in December

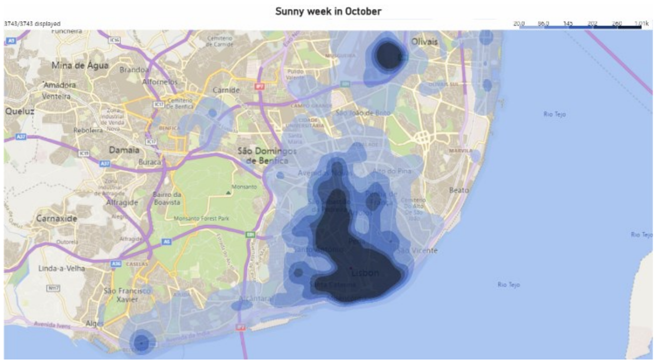

Fig 6 -Tourists' mobility on a sunny week in October

B) Case 2 – Web Summit

Web Summit is an annual technology conference held in Lisbon, Portugal and the biggest in Europe [21]. In 2021, still under the effects of the COVID-19 pandemic, the event gathered over 40,000 visitors. This influx of people allowed us to visualise the impact that the event had in the area where it was held (FIL – Lisbon International Fair).

To analyse this, we mapped the movement of Roaming users and considered those who were staying in the same grid for more than 5 minutes when the event was taking place and put those results side by side with the remaining days of the month. On the Web Summit period (1^{st} – 4^{th} November 2021), we were able to register an average of 2852 tourists at any given time in the event area. Compared to the 262 for the rest of the month of November (4^{th} – 30^{th} November 2021). This is a 987% increase in tourist activity (Figure 7 shows this example).

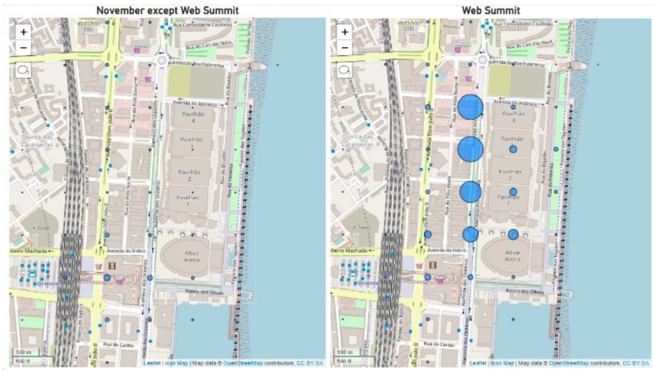

Fig 7 - Comparison of Tourist Mobility Before, and After vs. During Web Summit

C) Case 3 – Big Events Monitor Process – Case of a Game Day

Football tourists are those who travel to attend a football event, often a game. Within this group, it is feasible to identify three important categories of football tourists, the most common of which are travelling football teams.

The proposed approach allows for a real time monitoring of the tourist's movements in the city. Since the UEFA Champions League games create big movements of people from around the world. We monitored 4 major games in Portugal.

From the 4 games we monitored we chose the one which brought tourists from a nationality a lot less present throughout Lisbon, throughout the months in study, than the nationalities of other football fans that came to Portugal to see the other games (for privacy purposes, in this paper, we will call it nationality A). This aspect allowed us to monitor the football fans' movements a lot more effectively and clearly, than if we had a considerable number of tourists just visiting the city in general.

We were able to identify the time of arrival of the football fans at Lisbon Airport – 9 am – the day of the match, as can be seen in Figure 8 (big dot in the Humberto Delgado Airport).

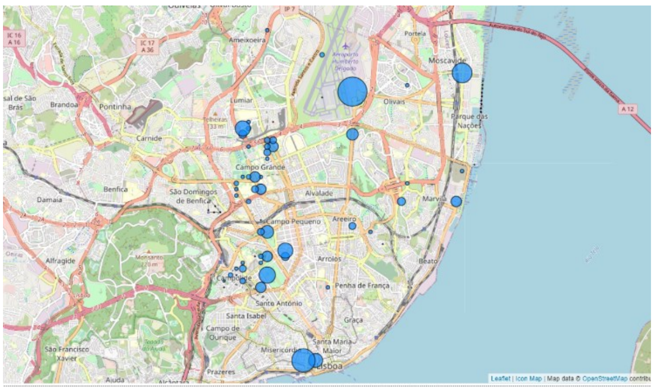

Fig 8 - Arriving time of football fans from nationality A in Lisbon Airport at around 9h

During the day they went mainly to the historic downtown area, as we can see in Fig 9.

Then at the game time we identify around 5000 tourists from nationality A at Sporting stadium (see Fig 10). After the game, since there are no flights, they stay and sleep in Historic Downtown Area (see Fig 11) and next day around 9h they leave Lisbon (see Fig 12). This case shows the importance of this study for parishes to understand tourists' movements and to better manage big events.

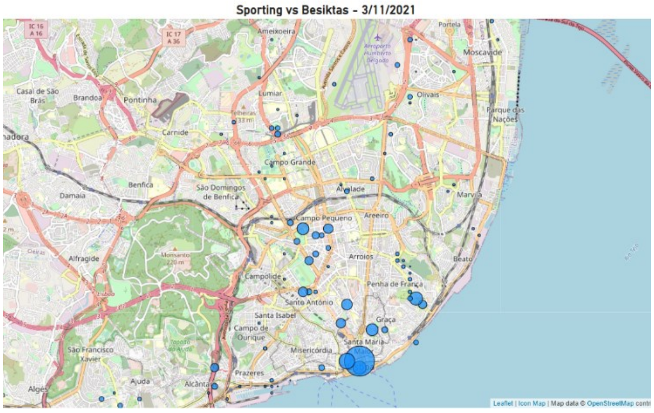

Fig 9 – Fans from nationality A at 14h in the Historic Downtown Area

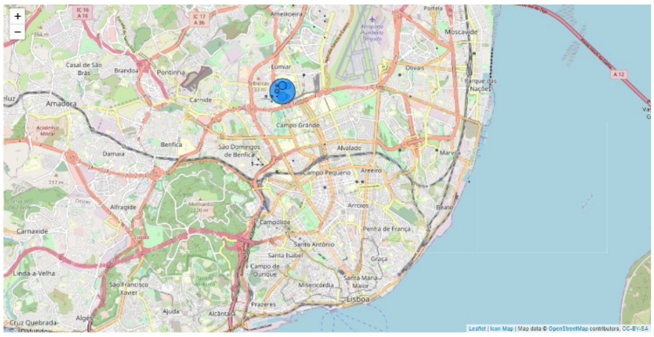

Fig 10 – Fans from nationality A during football match

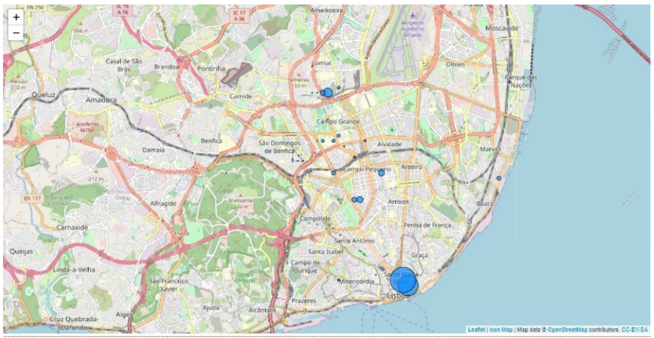

Fig 11 - Fans from nationality A after football match

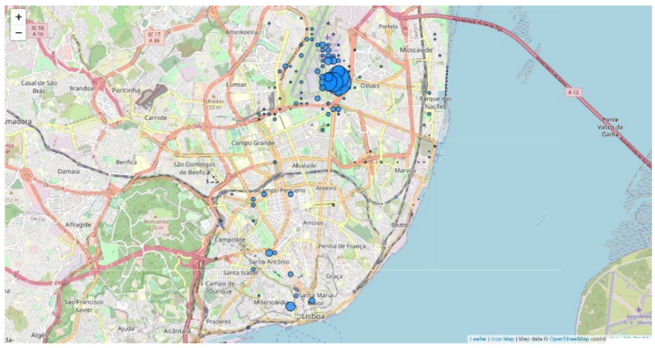

Fig 12 – Fans from nationality A the day after the match at 9h leaving Lisbon

A) Case 4 –Shopping activity

Shopping activity can be checked using the communication antennas that cover the shopping mall. The number of tourist visitors is useful information for authorities and store owners. Visit patterns and understanding abnormal behaviour during promotion days like black Friday is useful information. We can check where tourists shop more and see if these promotions also influence their behaviour. The data collected shows this influence and it is possible to witness this behaviour based on the nationality. Figs 13 and 14 show this influence and also the main place where tourists shop.

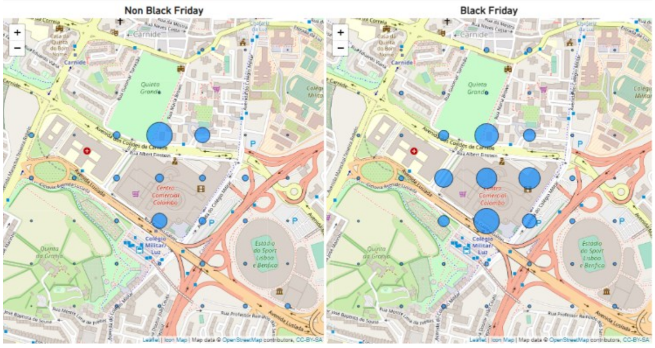

Fig 13 – Effect of Black Friday in Colombo Mall

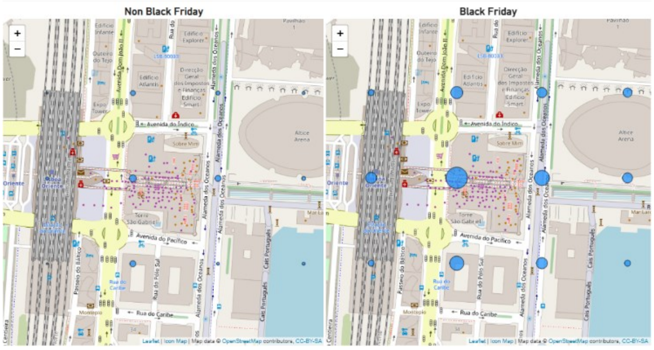

Fig 14 – Effect of Black Friday in Vasco da Gama Mall

V. CONCLUSIONS

With this study we were able to understand the mobility patterns of tourists in Lisbon.

They tend to frequent the historic downtown and centre area, and explore the coastal areas, which have points of interest.

We verified that after the 20th of each month there is usually a decrease in the presence of tourists in Lisbon.

The precipitation discourages tourists from visiting tourist sites. They tend to stay in the "hotspots" (the centre area and the historical downtown area), visiting museums, shopping malls, or staying in the comfort of their place of stay. When it's sunny, tourists tend to spread out more around the city and visit more areas than they stay in the hotspots (behaviour seen when it's raining).

And finally, international events bring a high number of tourists to the city where they're being held and especially to the surrounding area.

Understanding the geographical distribution of urban tourism is crucial for public policy. Thus, in areas with a large concentration of visitors, local authorities may consider initiatives to enhance the tourist experience, such as developing pedestrian-only lanes or enlarging sidewalks, expanding public places with free Wi-Fi, and situating new tourist information centres, among others. The big picture in real time provided by current developed work allows local authorities to understand this movement, identify dangerous concentration in big events.

We have developed a platform that collects anonymized data from mobile operators. The presence of mobile operators in the market allows for a representation of a city's population (if there are multiple operators in a city). Our approach started with a data cleaning and processing developed in python and then a visualization of the data in maps and dashboards made in PowerBI. This can be replicated in other cities and it's an important tool for city management authorities to understand the concentration and movements of tourists and adjust processes and facilities.

This work allows us to understand tourist movements and their patterns, correlated with weather, nationality and events. Modeling tourist behavior is another important topic.

ACKNOWLEDGMENT

We thank the Lisbon City Council for providing us with the data necessary for this study, namely Mr. António Costa (Lisbon City Council), Mrs. Helena Martins (Lisbon City Council) and Mrs. Paula Melicias (Lisbon City Council). We also thank IPMA for providing us with the data necessary to complement our study.

This work was supported by EEA Grants Blue Growth Programme (Call #5). Project PT-INNOVATION-0045 – Fish2Fork.

REFERENCES

[1] LCG, "Toursim numbers in Portugal | March 2022." https://travelbi.turismodeportugal.pt/turismo-em-portugal/turismo-em-numeros-marco-2022/ (accessed May 13, 2022).

[2] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement," *BMJ*, vol. 339, p. b2535, Jul. 2009, doi: 10.1136/bmj.b2535.

[3] M. Fekih, T. Bellemans, Z. Smoreda, P. Bonnel, A. Furno, and S. Galland, "A data-driven approach for origin–destination matrix construction from cellular network signalling data: a case study of Lyon region (France)," *Transportation*, vol. 48, no. 4, pp. 1671–1702, 2021, doi: 10.1007/s11116-020-10108-w.

[4] C. Badii, A. Difino, P. Nesi, I. Paoli, and M. Paolucci, "Classification of users' transportation modalities from mobiles in real operating conditions," *Multimed. Tools Appl.*, vol. 81, no. 1, pp. 115–140, 2022, doi: 10.1007/s11042-021-10993-y.

[5] E. Carter, P. Adam, D. Tsakis, S. Shaw, R. Watson, and P. Ryan, "Enhancing pedestrian mobility in Smart Cities using Big Data," *J. Manag. Anal.*, vol. 7, no. 2, pp. 173–188, 2020, doi: 10.1080/23270012.2020.1741039.

[6] J. Jeong *et al.*, "Mobility Prediction for 5G Core Networks," *IEEE Commun. Stand. Mag.*, vol. 5, no. 1, pp. 56–61, 2021, doi: 10.1109/MCOMSTD.001.2000046.

[7] X. Li *et al.*, "Position prediction system based on spatio-temporal regularity of object mobility," *Inf. Syst.*, vol. 75, pp. 43–55, 2018, doi: 10.1016/j.is.2018.02.004.

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

[9] P. Guo, K. Xiao, Z. Ye, and W. Zhu, "Route Optimization via Environment-Aware Deep Network and Reinforcement Learning," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 6, 2021, doi: 10.1145/3461645.

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

[11] C. Mizzi *et al.*, “Unraveling pedestrian mobility on a road network using ICTs data during great tourist events,” *EPJ Data Sci.*, vol. 7, no. 1, 2018, doi: 10.1140/epjds/s13688-018-0168-2.

[12] P. Keawsomnuk, “A structural equation model of factors relating to smart cities that affect the management of the world heritage site as well as the quality of life of tourists and villagers in Ayutthaya, Thailand,” *Humanit. Arts Soc. Sci. Stud.*, vol. 21, no. 1, pp. 35–42, 2021, doi: 10.14456/hasss.2021.4.

[13] A. Cheshmehzangi *et al.*, “A hierarchical study for urban statistical indicators on the prevalence of COVID-19 in Chinese city clusters based on multiple linear regression (MLR) and polynomial best subset regression (PBSR) analysis,” *Sci. Rep.*, vol. 12, no. 1, 2022, doi: 10.1038/s41598-022-05859-8.

[14] M. Šauer, J. Vystoupil, M. Novotná, and K. Widawski, “Central European Tourist Flows: Intraregional Patterns and Their Implications,” *Morav. Geogr. Rep.*, vol. 29, no. 4, pp. 278–291, 2021, doi: 10.2478/mgr-2021-0020.

[15] X. Lao, X. Deng, H. Gu, J. Yang, H. Yu, and Z. Xu, “Comparing Intercity Mobility Patterns among Different Holidays in China: a Big Data Analysis,” *Appl. Spat. Anal. Policy*, 2022, doi: 10.1007/s12061-021-09433-z.

[16] N. X. Leow and J. Krishnaswamy, “Smart cities need environmental consciousness and more social responsibilities as an outcome of COVID-19 – reflections from urban road commuters,” *Foresight*, vol. 24, no. 2, pp. 276–296, 2022, doi: 10.1108/FS-02-2021-0035.

[17] U. Türk, J. Östh, K. Kourtit, and P. Nijkamp, “The path of least resistance explaining tourist mobility patterns in destination areas using Airbnb data,” *J. Transp. Geogr.*, vol. 94, 2021, doi: 10.1016/j.jtrangeo.2021.103130.

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

[19] R. Wirth and J. Hipp, “CRISP-DM: Towards a Standard Process Model for Data Mining,” p. 11.

[20] K. Healy, *Data Visualisation: A Practical Introduction*. Princeton University Press, 2018.

[21] “Web Summit: Lisbon hosting the largest tech event in the world - Portugal - Portuguese American Journal.” https://portuguese-american-journal.com/web-summit-lisbon-hosting-the-largest-tech-event-in-the-world-portugal/ (accessed May 13, 2022).