NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

Cities Mobility Management

Mobility prediction model applied to Lisbon marathons

Fábio Miguel Domingues

Project Work proposal presented as partial requirement for obtaining the Master's degree in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

CITIES MOBILITY MANAGEMENT

by

Fábio Miguel Domingues

Project Work proposal presented as partial requirement to obtain the Master's degree in Information Management, with a specialization in Knowledge Management and Business Intelligence

Co-supervisor: Miguel de Castro Neto

Co-supervisor: Pedro Alexandre Reis Sarmento

Co-supervisor: Marcel Motta do Nascimento

May/2021

DECLARATION OF ORIGINALITY

I declare that the work described in this document is my own and not from someone else. All the assistance I have received from other people is duly acknowledged and all the sources (published or not published) are referenced. This work has not been previously evaluated or submitted to NOVA Information Management School or elsewhere.

Lisbon, 12/05/2021

Fábio Miguel Domingues

(the signed original has been archived by the NOVA IMS services)

ACKNOWLEDGEMENTS

To complete my master thesis, I have received a great support and assistance.

First, I would like to thank all my co-supervisors, Miguel de Castro Neto, Pedro Alexandre Reis Sarmento and Marcel Motta do Nascimento. Three distinct personalities with different expertise's that were crucial in the development of my entire project by helping me overwhelm complex technical and functional topics, pushing me further to overcome my concerns and limits in multiple viewpoints.

I also would like to thank all my colleagues, with a special thanks to those who worked close to me (smart cities projects team) that also helped me in my process of acquiring new skills that made this master thesis development possible.

Finally, I would like to thank my entire family, the ultimate ground support giving me wise counsels, sympathetic ear and patience when helping me right down all my process in a clearer way. Besides providing me true joy and happiness to my life outside of my research.

ABSTRACT

Lisbon, as Portugal capital, is a city that receives a vast number of events every year. These enormous concentrations of people, in the same place, at the same time, requires huge transport services planning and organization. To elevate Lisbon to a state of what is called smart city, smart mobility topic must be improved, in a serious way. To achieve this, Lisbon must be able to stop being reactive and start being proactive. Predicting people's behavior, concerning to the theme "mobility" is something that can improve drastically population's life quality. With this study, it is intended to have a better and specific understanding of how CARRIS public transport service is managed in Lisbon, during city's marathons. The main objective is to implement smarter mobility strategies, during big events, analyze people behavior before, during and after these events and try to predict future population behavior, based on data.

CARRIS and other related sources provided data that was integrated into a database system in an automated way. Above this database system a machine learning prediction model was put in place revealing that it is possible to forecast at least 75% of the attendance on this type of transports. This project will end with Power BI self-explanatory reports that can help decision making. By having a better understanding of all problems related to smart mobility during big events, Lisbon may predict and act accordingly. By improving smart mobility, the city is improving life quality, as well.

KEYWORDS

Smart mobility; Smart Cities; Smart event planning; Citizen's life quality; massive mobility

ACRONYMS

AML Área Metropolitana de Lisboa

ICTs Information and Communication Technologies

INDEX

Figures Index

Tables Index

1. INTRODUCTION

1.1. BACKGROUND AND PROBLEM DEFINITION

According to a review concept "smart city is multi-faceted". Descriptions of smart cities are now including qualities of people and communities as well as ICTs" (Albino, Berardi & Dangelico, 2015)."A Smart City is therefore a complex, long-term vision of a better urban area, aiming at reducing its environmental footprint and at creating better quality of life for citizens. Mobility is one of the most difficult topics to face in metropolitan large areas. It involves both environmental and economic aspects, and needs both, high technologies and virtuous people behaviors" (Benevolo, Dameri, & Auria, 2016).

Portugal has become a first option when choosing a stage for major events, for many reasons such as the mild weather and historical environment. Lisbon, in concrete, receives world events like *WebSummit,* which received 70 thousand people (Barbosa, 2018). In addition to this major event, Lisbon have many other big events like *Rock in Rio*, *Santa Casa Alfama, EuroVision, Festas de Lisboa, Sumol Summer fest, ModaLisboa – Lisboa Fashion Week, Meia Maratona Lisboa, Football Classic Matches.* However, it's not everything great and the city has "a long journey to become a smart city" (Lopes, 2018).

"Lisbon is the Iberian city with more traffic, even though it's among the best in Europe in mobility" (David, 2017). Our city Portugal has 10.3 million people, where 2 821 876 are Lisbon Metropolitan Area citizens (AML, 2020). In international big events, Lisbon should never forget to include in its account, foreigners (from other Portugal cities and other countries). For the purpose of establishing urban and regional innovation ecosystems, Lisbon must have an environment turned to share common resources, sustainable partnerships and cooperation strategies, among the main stakeholders (Schaffers, Komninos, Pallot, & Trousse, 2011).

Usually, there is a lack of public transports to serve so many people, at the same time. When major events take place at rush hours, the problem gets even worse. To face these massive movements of people inside Lisbon, there is a need to predict all these movements and plan accordingly. This way, Lisbon will be able to ensure enough transport capacity and quality service to its population, including possible foreigners, where it is most needed. Optimize public transport service during major events will also make Lisbon a first choice to international events.

Some studies, regarding smart mobility in metropolitan area of Lisbon, have already been made, analyzing people's behaviors. However, data collected did not focus on big events migration prediction. Undertaking this problem, this project will be able to retrieve better conclusions of our population behavior during big events and help plan appropriately, forecasting citizens migration using data, with a higher level of confidence, taking in consideration Lisbon citizens major needs and apprehensions. Hence, to accomplish this, "especially in the most mature phases of the Smart Mobility implementation, each citizen must be a proactive actor, by accepting a limitation in its own transfer freedom (reducing the use of private car, for example) and embracing the pursuing of shared smart aims "(Benevolo et al., 2016).

1.2. STUDY OBJECTIVES

The creation of a data structure that helps control and optimize Lisbon public transports strategy during marathon days (big events) is the main goal of this project.

In general, this project intends to offer a better mobility service in Lisbon, focusing on big event's needs, by developing an analytic in-depth study of our population. This project wants to create dynamic reports based on a Datawarehouse that joins data from different sources. With these reports we expect to find patterns and help public transports decision making. Thus, this project can extract accurate knowledge and truly improves Lisbon citizen's mobility, contributing for a smarter city. It plans to help mobility services, as possible, and optimize theirs offers, in all perspectives.

This study will help public transport entities by offering them the opportunity of having a better understanding of their end user essential concerns. Act according to knowledge retrieved from data instead of intuition, start being proactive in place of reactive, having a smarter mobility system and consequently a smart city environment, improving Lisbon reputation in terms of organization provided during big events.

1.3. STUDY RELEVANCE AND IMPORTANCE

By gathering all this data about major events mobility, across multiple public transport institutions, this project may answer these main questions:

1. What are the Lisbon citizen's behavior during these events?
    - Where are the most requested bus stops and routes?
    - When are those bus stops and routes more requested?
2. What would be the solution that would give people better quality of life regarding smart mobility and consequently generate a smart city environment during big events?

This project is taking a step forward in terms of smart mobility, basing our plan in data analyses. Moreover, Lisbon will have a chance to plan the future of multimodal mobility during future events enriching city's mobility system, by reaching a better understanding of our population choices regarding their migration patterns.

Planning smart mobility is important in a sense that it will improve quality of people's lives and provide a better leisure experience. This project will also help public transport entities optimize their services and consequently, rise their profits.

Event planner may also make use of this development to achieve better results regarding their attendee's satisfaction.

Power BI will be used to explore and present information in a more interesting way, in a perspective of an event planner, transports operator planner and a crowd manager. Having a platform that put together all these perspectives creating the opportunity to plan and act accordingly to citizens needs regarding mobility. Having a city that can auto organize itself even when having uncommon behavior is an upgrade of life quality and consequently what is also called smart mobility.

In conclusion, I will analyze all the data and look for patterns in data that may answer the questions mentioned in this study, to predict future event necessities and enlarge Lisbon has a smarter event planner, beyond borders.

1.4. STUDY DESIGN AND STRUCTURE

This study follows the normal course of a master's thesis project research, using a predefined empirical methodology that instead of being centered in a single identity it is mix-centered as it tries to assist multiple entities as an overall.

The thesis is divided in seven chapters. It starts with the - Introduction – where the most relevant arguments that will be discussed are announced, to inform readers of its importance, context, and objective. On the second section – Literature Review – is given to analyze the current situation, study background and other projects recommendations. The third section – Methodology – is where the approach applied is presented and its applicability is justified according to the project scope. The fourth segment is – Conceptual Model – that focus is efforts on developing an in-depth exploratory analysis in order to support the final prediction model, according to the overall environment. The fifth stage is – Prediction Model Development – where the final model is built to address the predefined problem and reclaim its respective answers. The sixth chapter – Conclusion – points out the most relevant assumptions retrieved from this project materials at the same point that explicitly answers this master's thesis target questions. The seventh and last chapter – Limitations and Recommendations – uncovers all the existent project constraints unsolved. At the same time, it proposes future study pathways to improve and boost beyond the current progress.

2. LITERATURE REVIEW

2.1. SMART MOBILITY DURING BIG EVENTS

To have a better understanding regarding this theme, first of all, Lisbon should look at the topic as 3 different blocks and then form a connection between them, since they all have a layer that overlaps between them:

- Crowd Management
- Event planning management
- Public transports management

By linking these 3 puzzle pieces, Lisbon may have better insights on how it is possible to upgrade life quality regarding smart cities mobility, in special big events.

I will review all the work done regarding smart mobility, in order to be able to create a structure that help Lisbon services provide a better solution as an all. Joining public transports and event planner entities, having as the main goal, mobility improvement and consequent life quality improvement.

Public transports services are still not working as one. Besides this fact, events planner entities are not always planning close to public transports operators, in a synchronized perspective basis.

My objective does not focus only on offering public transports operators a way to monitor and start creating a model to predict crowd movement, but also start developing a structure that looks into all these, as one major problem. These crucial problems may be solved with one solution, based on unification and information centralization.

Weather is also an important fact that can be taken in consideration regarding smart mobility planning and prediction. A rainy day is obviously a factor that changes people's choices regarding transportation. Having too cold or too hot days, is something that can also be predicted and used to.

Many other possible aspects may influence crowd movement and choices regarding transportation. However, for this particular work, the focus is on some of these points mentioned above. It intends to give a first step towards smart mobility data centralization between major entities involved by being able to exploit knowledge that benefits all Lisbon Citizens.

As referred by Neto and Cartaxo (2020), cities are par excellence places where big data thrives, given that it is produced by the Internet of everything (as interconnected systems, sensors and people) what creates an opportunity to politicians and professionals to answer cities day-to-day challenges, such as more efficiency in services and infrastructures management, increased quality of life, and addressing the problems of climate change taking advantage of information management and data science capabilities to measure and describe the realities that happen in the territory of the city. Furthermore, with these solutions, it is possible to predict what happens (very often in real time) and then take rapid decisions to prescribe what could be the actions and the following course of actions. Policy making can, consequently, more easily adapt to fact-based urban environments.

2.2. CROWD MANAGEMENT

The study of real time crowd movement has begun more than 20 years ago. When studying mobility flows, Lisbon must take in consideration that there are tons of difference facts that influence people behavior. However, there are some layers that overlaps crowd's behavior, "the behaviour of a crowd as a whole is often highly organized and certain spatial-temporal patterns appear at a macro scale" (Heldens, Litvak, Steen, & Senior, 2018). For example, the "type of crowd along with external factors, such as the weather, have an impact on both density, flow and walking speed"(Larsson & Ranudd, 2019).

Events are part of a booming industry that continues to grow both domestically and internationally and all this popularity leads to even more and more attendees, making crowd management and its planning absolutely necessary, in order to have a controlled and safe environment to display any event. According to (Anna, Abbott, & Geddie, 2001), a successful crowd control plan first includes a statement of purpose that focuses the plan and provides for crowd control goals.

On (Leung, Chan, Hui, & Li, 2011) analyses over San Francisco and Shanghai networks and traffic flows there in measured by GPS taxis lines it was concluded these techniques used solely is not enough to predict mobility. There are many factors that bias this approach, starting with the fact that people do not always follow principal roads, sometimes they use secondary routes. Population's culture is also a feature that influences population's choices regarding mobility. Using GPS to better understand mobility is a good way to gain better insights, however these notions should not be used solely but combined with other types of relevant insights.

Crowd behaviour is not always linear in a sense that sometimes there are other external factors that may cause unexpected behaviours, like a wild rush to force entrance or exists in an event, rush while driving a car increasing the probability of car accidents, anxiety and completion when there is the possibility of losing a promotional event, scramble to get event tickets. All these factors influence crowd movement and must always be an alert that crow management is not always a linear analyses.

Taking in consideration (Khozium, Abuarafah, & AbdRabou, 2012), a large moving crowd has the capacity to 'self-organize' safely if density is low enough. Under normal conditions, crowds have a spontaneous intelligence of their own, developing streams, that keeps everyone moving. As crowd density rose, they identified the onset of stop-and-go waves like those found in road traffic jams. This was followed by transition to a much more chaotic state, with outbreaks of panic as individuals lost control. This kind of occurrence known as crowd turbulence can culminate into disasters.

The intrinsic relationship between the two spatial logics (separations and circulations), need to be empathised. The two logics complement each other; indeed, the example of mega-event security powerfully illustrates the intertwined logics of fixing, enclosing and delimiting space on the one hand, and off managing, guaranteeing, and improving circulations on the other. However, in the event's context of increased density, coexistence and risk exposure, the core requirements for the management of enclosures and circulation, fixity and fluidity also compete with each other, exclaims (Klauser, 2012). In other words, Lisbon must investigate this matter from two points of view that must be balanced. The first one is space and freedom in a sense that people are able to manage themselves without intervention in most cases, however, when speaking about huge events, where the risk of something going wrong is higher and a small problem may trigger of something catastrophic. Lisbon

must force the creation of boundaries and common paths, just for precaution and coverage that crowds stick with their path safely and peacefully.

2.3. EVENT PLANNING MANAGEMENT

Event planning is crucial since as events grow in popularity, attendances also increase. Due to this rapid increase, crowd management and crowd control are now important issues in this industry". (Anna et al., 2001).

To retrieve knowledge from this data, attendance data and event location will be retrieved. Furthermore, it will be possible to join this data with all the other indicators of crowd movement in the same spatial-temporal environment.

Pre-event planning should begin 12-18 months before the date of the event, if possible (Connors, 2007). However, at federal level, pre-event planning may start even earlier. There are other key partners that should be included in big event planning like fire departments, emergency medical services (EMS), transportation, public works, health, other public agencies and the private sector—businesses affected by the event, as well as private security.

Taking in consideration (Ewen, 2019) review concept there are some big opportunities regarding data driving event industry.

- **Improving Targeted Promotions**

Events require many monetary expenses. Based on data, it is possible to evaluate our population and optimize resource, maximizing audience outreach.

"Location allows event marketers to connect their digital marketing efforts to how prospective attendees behave in the real-world. As a result, event marketers can provide even more personal advertising to their target audience" (Ewen, 2019).By having access to this information event manager will not only be able to attract more attendees, but also personalize event experiences their guests.

Data gives a closer insight about attendees, a touch that event managers can use as leverage when developing an event plan.

- **Gaining Insight from Analytics**

Data gives a closer insight about attendees that event managers can use as advantage when performing an event plan.

Event planner will be able to predict trends. Predictive analytics are an optimal way to improve by making more near-accurate choices regarding event planning. (Ewen, 2019)

- **Personalizing Attendee Experiences**

After having attendee's insight, it is possible to personalize experiences, elevating attendee's experience. Give transport offer advices or even free tickets, for example. (Ewen, 2019)

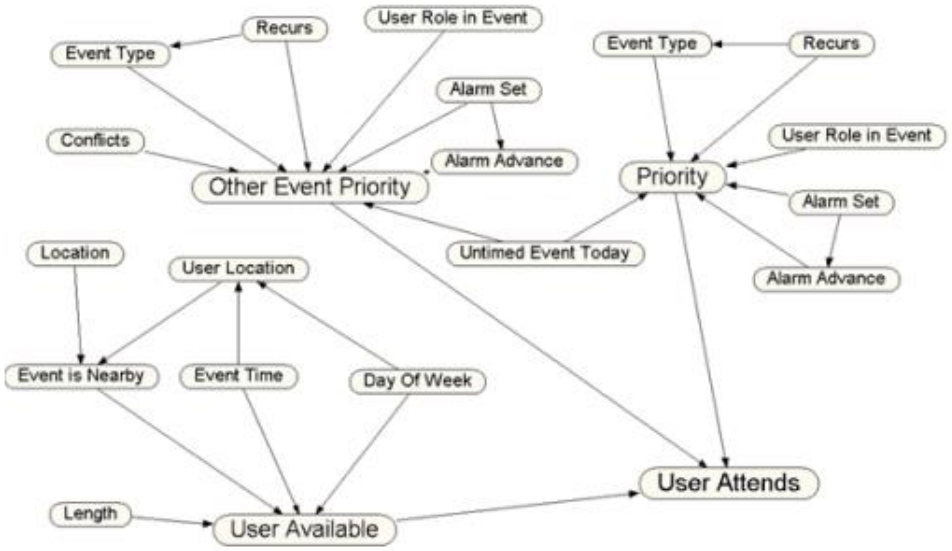

Figure 1 – A Bayesian model of event attendance (Mynatt & Tullio, 2001)

"Though the task of predicting which events on a user's calendar will be attended seems infeasible, Lisbon can make use of attendance history and attributes of calendar events to present a likelihood of attendance to applications" (Mynatt & Tullio, 2001).

In addition to all this, you can use data from previous events to improve upcoming events.

2.4. PUBLIC TRANSPORTS MANAGEMENT

Figure below reveals that Lisbon municipality is not even on top 10 smart cities of Portugal. As a capital, this is a problematic theme that clearly should be addressed.

RANKING	MUNICIPALITY
1	Porto
2	Águeda
3	Cascais
4	Bragança
5	Guimarães
6	Matosinhos
7	Braga
8	Sintra
9	Aveiro
10	Santarém

Figure 2 - Top 10 smarter municipalities in Portugal (Lopes, 2018).

Considering (Rao & Rao, 2012) review "there are two principal categories of causes of congestion, and they are; (a) micro-level factors (e.g. relate to traffic on the road) and macro-level factors that relate

to overall demand for road use. Congestion is 'triggered' at the 'micro' level (e.g. on the road), and 'driven' at the 'macro' level by factors that contribute to the incidence of congestion and its severity". Where micro level factors are the fact that many people want to move at the same time, too many vehicles for limited road space. In the presence of events that are irregular, trips may be delayed (accidents, vehicle breakdowns, poorly timed traffic signals, special events like mass social gatherings, bad weather conditions, and many others). On the other hand, macro level factors are land-use patterns, employment patterns, income levels, car ownership trends, infrastructure investment, and regional economic dynamics, many others). All previous factors are as much critical on a daily perspective as they are during an event planning.

(Rao & Rao, 2012) also concluded that average speed, flow/density, delay and travel time variability could all be used to assess the level of congestion.

To better define congestion, it must exist threshold values for the beginning of what common sense call delay. Threshold definition is crucial to define consequences (improvement actions) on their trespass.

There are some cases where the approach of analysing congestion with a threshold based on capacity is also a good strategy. These applies to most transport services (metro, bus, train). Creating a co-threshold could be the optimal approach on some cases.

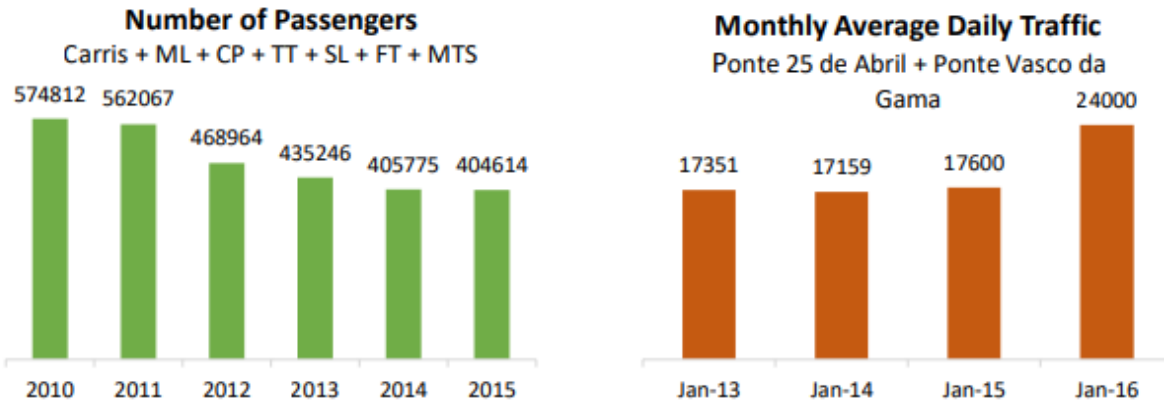

Figure 4 - Number of passengers 2010 –2015 of the seven major transportation operators & Monthly average daily traffic on the two bridges in the Lisbon municipality (Lopes, 2018).

According to (Lopes, 2018), number of passengers using major transportation operators in Lisbon has decreased between 2010 – 2015. On the other hand, traffic on the 2 bridges (Ponte 25 de Abril & Ponte Vasco da Gama) has increased in more or less the same period. At that time. it was concluded that people tended to stop using public transports and start using their personal vehicle. This is something very common when the population is not satisfied with the existing offer. However, on 26 march of 2019, the government has announced a new transports policy that have affected positively those major transports operators by decreasing multimodal pass price (Redação, 2019). Passenger traffic on Portugal public transports has increased by 8% in the second quarter on all modes of transport year-on-year. It seems that Lisbon citizens are giving a second chance to public transports entities (Lusa & Público, 2019). But with the increase of the demand there is also the need to increase the offer in a smart and sustainable way, and this way should be based on data, facts over intuition.

According to (Daraio et al., 2016), the matrix definition of transports network is the key to identify population wishes regarding there mobility, by establishing direct routes where there is more demand. This hierarchy structure definition can start by creating a primary network that joins major crowd congestions spots (hot points). This primary network should be complemented with secondary connections that covers the entire urban area in question. These secondary routes should focus more on demand concerns. However, a strongly hierarchical network approach must be taken, with caution, since it also increases the number of transport change which gives passenger and higher level of discomfort.

	Length (m)	Width (m)	Height (m)	Capacity
Standard	11.5	2.5	3	90
Articulated bus	17.5	2.5	3	135
Minibus	7	2	2.25	20

Table 1 – Public bus features (Pires da Costa, 2008)

Concentration (pas/m2)	Area/pas (m2/pas)	Commodity and mobility quality
<1	>1	No contact. Easy movement.
2 or 3	0.50/0.33	Small contact. Movement with some discomfort
4	0.25	High contact. Hard movement.
5	0.2	Permanent contact. Very hard movement
>6	0.15	Strong contact. Forced movement.

Table 2 – Public bus commodity and mobility quality (Pires da Costa, 2008)

Considering (Pires da Costa, 2008), public transports offer out of rush hours is normally defined as the minimum fixed interval for the type of service in question. Even though, 10 minutes is the normal threshold for the maximum time that a person is willing to wait for a transport, there are situations where the demand may not justify high frequencies. This type of system requires regular demand. As an alternative, instead of high frequency of a transport offer, a strategy can be implemented based on schedule, where punctuality is crucial to make this option reliable.

On rush hours public transport, the offer must be higher, combining vehicles higher capacity and higher frequencies. The offer must be higher than demand, with a safety margin, based on the capacity utilization coefficient (or load factor). This load factor mainly depends on passenger's level of comfort and operation costs.

The type of passengers also influences the type of demand, if Lisbon have mainly elder passengers it probably needs more seating, the opposite happens with young students. The city must be aware that a higher percentage of seats, against standing places, offers less flexibility to transport more passengers.

Transport Method Capacity (F = Full & M = Moderate)	Interval between vehicles (sec)	Line Capacity (car./h), c = f * n (car./h)	Vehicle Capacity (car./h), c=f * n (car./h)	Max Capacity C (places/h)	Relative Capacity vs relative capacity of standard bus	Operation velocity (km/h)	Productive Capacity Pc (capacity x Km per h)
Standart bus F	30	120	75	6 000 - 9 000	1.0	8 - 12	75 - 50
Standart bus M	40	90	75	4 000 - 6 300	1.0	20 - 40	220 - 160
Articulated bus F	33	110	110	8 500 - 12 000	1.4	7 - 11	90 - 60
Articulated bus M	45	80	110	5 000 - 8 500	1.3	18 - 36	290 - 180
Metro F	100	36*10=360	175	40 000 - 63 000	6.9	22 - 40	1596
Metro M	120	30*8=240	175	30 000 - 42 000	7.0	25 - 45	1260

Table 3 – Public Transports Features Comparison (Pires da Costa, 2008).

2.5. MOBILITY MEASUREMENT MEASURES AND TECHNIQUES

Data retrieved from telecommunications services, Cell Towers, it is a useful way to extract some know how regarding people concentrations.

Data Drawbacks

- Cell Tower do not give us the exact location of every device, they only give us an action ray (coverage).
- Devices can be signed in multiple towers at the same time.

Besides drawback of this kind of data, it is possible to infer some crowd behaviours when join this type of data with data from public transports and events.

Technology	Data	Reference	Accuracy	Coverage
Cell Tower	Cell owner ID + signal strength or geographic coordinate	Relative/Absolute	50-200 meters in cities	Cell coverage. 5-30km from cell tower

Table 4 - Cell Tower features (Pan, Qi, Zhang, Li, & Wu, 2013)

The error rate is high when deducing device location area based only on one tower, however, if this project uses centroids of two or more cell towers, diminish error becomes a possibility.

Overall, by extracting data from cell towers, Lisbon can have an idea of people's concentrations in an action ray. Lisbon expect to get a higher volume of people close to those places where it is planned to

have a big event. The city will be able to monitor crowd's movement, by Joining this data with public transports data.

According to (Addison & Heydecker, 2015) traffic flow analyses are upon to three quantities of speed (v), flow (q) and density (k). These three measures lead us to the following formula:

- $q = kv$

Where:

- **q** is flow, measured in vehicles per unit of time.
- **k** is density, measured in vehicles per length of road.
- **v** is speed measured in length per unit of time.

However, these measures must be used with, caution because there are some external factors that bias their results:

- It does not take in consideration that vehicles have different lengths.
- Roads have different length and with width.
- Traffic regulation can also influence.

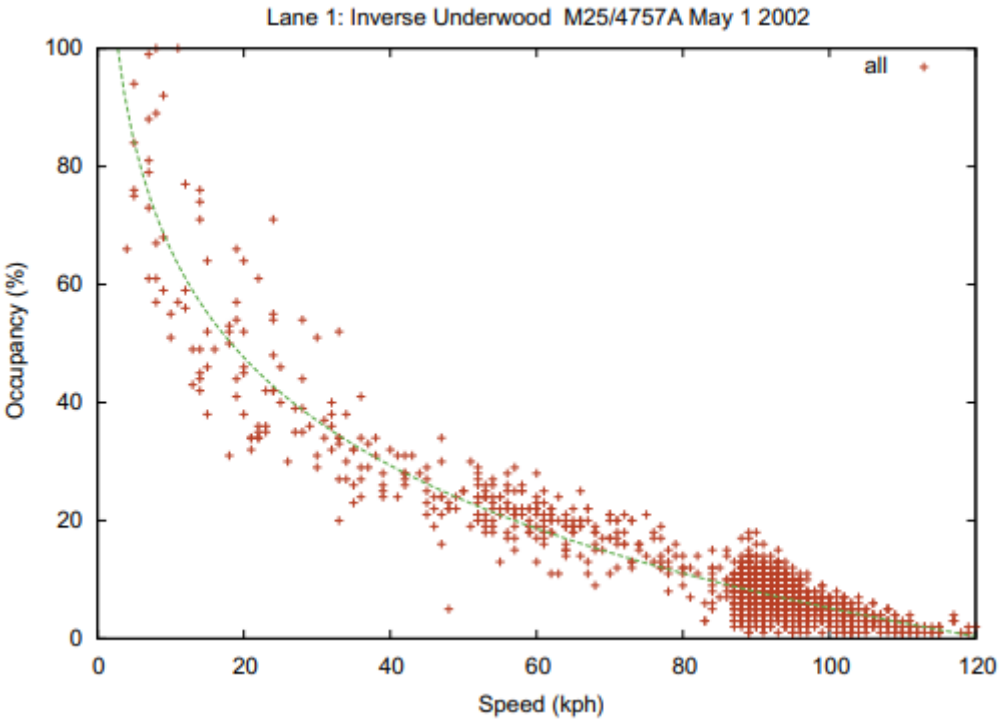

Figure 3 - Observations of Occupancy and Speed (Addison & Heydecker, 2015).

Regarding public transports effectiveness, there are many variables that can be used to evaluate performance, depending on the type of public transport. According to (Daraio et al., 2016), there are variables that try to reveal the performance from the producer perspective and other variables that

look into the community point of view. Some of the more suitable and common measures used to evaluate performance, focusing on the objective of this project are described in the table below.

Basic Measures	Calculated Measures
Number of vehicles	Travelled kilometers * hours of work
Seat Capacity (total seats of the fleet)	Number of vehicles * travelled kilometers
Number of passengers	Seats capacity * travelled kilometers
Number of bus traffic trips on routes	Vehicles * hours of operations
Hours of operations	Seats capacity * hours of operations
Travelled kilometers	Passengers * travelled kilometers

Table 5 – Performance Measures According to (Sampaio, Neto, & Sampaio, 2005)

Considering (Sampaio et al., 2005) statement, some mobility notions and possible measures to evaluate service quality of public transports are:

- The distance between user's origin and the initial station, as well as between the last station and the final station, can be used as system accessibility measure. "The shorter this 3 distance higher is route availability and, as a consequence, geographical coverage increases, making it easy and better to people to move from one place to the other." (Sampaio et al., 2005)
- Velocity joined with characteristics of the route can determine travel time. Nevertheless, Lisbon should always take in consideration traffic conditions and road quality too.
- Trustworthiness is also a factor that can be defined, as the lever of uncertainty of time schedules. It can be measured by comparing the number of trips on type with the delayed ones (delay). Since, "Punctuality bring users trust and fidelity." (Sampaio et al., 2005)
- Time intervals between each trip (frequency). Users appreciate knowing all possible changes along the day, weekday or any other type of special occasions.
- Seats capacity (maximum load), determined by the number of passengers in rush hour compared with the vehicle capacity.
- Vehicle features, as age, conservation, technology and all existing bringing's that influence passenger's comfort. Ex: Noise, temperature control.
- Existing information facilities on stations, such as schedules, timetable information and clear indications of stations and vehicles.

The spatial definition of a bus career is done through the identification of the origin, destination, its code, or id, route and respective distance. The location of all its stops and terminals should also be known (Pires da Costa, 2008).

The most used public transports spatial coverage indicators according to (Pires da Costa, 2008) are:

- Length of the line corresponding to the length measured in a direction along the line between its ends
- Total length of the axial network (or length of the network) given by the total length of the road network covered by public transport
- Spatial coverage rate given by the quotient between the extension of the network and the geographical area of a given entity (Ex: Parish, Municipality).
- Longitudinal coverage index, reflecting the part of the road network covered by the public transport network
- Population coverage rate given by the quotient between the served area and the population of a given entity Ex: Parish, Municipality)

The area served by a public transport can be defined by circles, centered on public transports stops with a radius of 300, or 600-meters equivalent to maximum walking paths of about 5 minutes, or 10 minutes, respectively. Although, Lisbon should be aware that, on urban area the real distance between two points on foot, is much higher since those routes are almost never straight. On these cases, it is preferable to use squares rather than circles. Those squares should also be centered on the stop and aligned according to its respective road, with a width of 500 meters, which is 250 meters for each side of the plan. This layout should have a maximum size of 1000 meters. If this project wants to apply an even more sophisticated study regarding the area of influence of the career, it should develop isochrones for each stop, lines that indicate all locations that are 5 or 10 minutes distant. This distance could also consider the slope of the route and apply a correction. Studies have revealed that a direct route of 400 meters is actually equal to 528 meters. When applying the same logic to a wheelchair person this distance increases to 930 meters.

Regarding public transports indicators of time coverage, they are normally based on operation period or pass-through frequencies (frequencies between passages is also a valid option).

Some examples of time coverage indicators are:

- Operating range
- Average frequency (average interval time between transports)
    - This indicator may change along the day or weekday.

Some possible indicators can also be made by joining special coverage with time coverage concepts, taking in consideration the existing offer:

- The quotient between total length traveled by vehicles, expressed as vehicles * km, and the area (or population) of a given geographical entity (ex: municipality). All this for a specified period.
    - Instead of using number of vehicles, Lisbon can also use the number of available places.
- Maximum number of entities that can pass through a location (ex: metro stop) per unit of time.
    - These entities can be carriages, vehicles or places.
    -

Other possible indicators are:

- Number of lines
- Network length
- Size of the fleet classified by vehicle type
- Number of stops / stations reflecting the spatial coverage of the service (accessibility)
- Percentage of passengers or trips made during a year
- Intensity of the service offered (the division of daily transport production by the network length)
- Average vehicles velocity.

- The quotient between the number of passengers and the distance traveled by vehicles within one year (higher values represent higher efficiency on an economist perspective).

Summing all points, mobility must be "in accordance with necessities, that is, routes should be planned to cover the whole area and allow flexibility in choosing an appropriate route. In addition, adaptations are required to attend passengers with motion restrictions. Besides quality requirements, efficiency is related to performance indicators, such as, low operational cost to users, minimum number of vehicles and personnel, but without a decrease in the quality of service provided. And efficacy is related to the number of users of public transport in relation to population, kilometers of routes provided in relation to area, and the satisfaction level" (Sampaio et al., 2005).

2.6. SMART MOBILITY PREDICTION MODELS

This project is trying to forecast the short-term bus passenger flow during marathons, acknowledging that, similar studies made on other cities are used as input to better understand what the best practices are to build a prediction model in the specific case of Lisbon Marathons.

Model	Variables	Parameters	Techniques	Model Evaluation	Model Comparison	Author
1 - IMM Algorithm (Time series model)	Passenger demand (15min, daily, weekly and hybrid)	SARIMA (p,d,q)(P,D,Q)^S Parameters chosen: (2,0,3),(1,0,0)^24	Stationary test was conducted by checking both trailing and truncation features of autocorrelation function (ACF), partial autocorrelation function (PACF), and augmented Dickey-Fuller (ADF) index.	MAE, RMSE, MAPE, and VAPE	AR, SARIMA ARMA.ARIMA and ARIMA-GARCH	Xue, Sun, & Chen, 2015
2 - MSRBF network 1. One-step-ahead prediction 2. Two-step-ahead prediction	Passenger demand on station inside time lag; Number terms selected changed between 10 and 12	Number of terms depends on GCV criteria.	1. NARX model 2. Fast fuzzy c-means 3. MPOLS algorithm and cross-validation (GCV)	RMSE, MAPE and VAPE	SVM, BRT, SSRBF	Li, Wang, Sun, Ma, & Lu, 2017
3 - Multilayer perceptron	The historical passenger flow data, the extracted EMD components and temporal factors are taken as inputs to build IMF's	Learning Rate: 0.2; Momentum: 0.8; Initial weights: 0.3; Hidden Layers: 2	Hybrid EMD–BPN	MAPE and VAPE	SARIMA	Wei & Chen, 2012
4 - LSTM	Historical passenger flow	Loops = 5; Learning rate -> Nadam: 0.002 SGD: 0.05 Rate decay: 0.9	LSTM SGD, LSTM Adagrad, LSTM RMSProp, LSTM Adam, LSTM Hybrid, LSTM SGD and LSTM Nadam Algorithm	MSE, MAPE and RMSE	Naive, ARIMA, SVR, LSTM SGD, LSTM Adagrad, LSTM RMSProp, LSTM Adam, LSTM Hybrid, LSTM SGD and LSTM Nadam Algorithm	Han et al., 2019

5 – Neural Networks using SCATS	Historical passenger flow; Traffic flow data (SCATS); weather; temporal variables: (ex: day-of-week, time-of-day)) , binary variables for peak hours	Hidden neurons: 7 neuros, 5 neuros, 7 neurons and 7 neurons	Neural networks using SCATS data (traffic flow data).	Analysis Of Variance (ANOVA), MAPE	Traditional models, Historical traffic flow model	Mazloumi, Currie, Rose, & Sarvi, 2009
6 - Demand Responsive Transport;	Tertiary student; 11-30 age groups; Education and work trips; Single person household; Low household vehicle ownership (0 or 1)	N.A	Correlation between possible variables	'Rank Order Centroid (ROC)'; 'Ratio Method'; or 'Pairwise Comparison'; R^2	N.A	Jain, Ronald, Thompson, & Winter, 2017
7 - Improved Spatio-Temporal Residual Networks	card id, bus route id, bus vehicle code, boarding and alighting time, latitude and longitude coordinates of boarding and alighting stations	Divided the region in 32 × 32 grids. Each grid had 0.625 km × 0.625 km , and the time interval is set to half an hour. The learning rate: 0.0001	Night hours where filtered from the dataset since the frequency of buses is much lesser during these hours. Algorithm is divided in two blocks: Spatio-temporal and scenario patterns.	RMSE and MAE	Historical average (HA), Autoregressive integrated moving average (ARIMA), DeepST, ST-ResNet	Liu, Zhang, Kong, & Yin, 2019

1. The core idea of the IMM algorithm is to estimate the prior optimized combinations of prediction models, dynamically using real-time data and filtering the predictions. The hybrid model makes effective use of the historical and real time information, showing a better explanatory power than the alternatives.
2. Regarding MSRBF, it was concluded that his network model can generate optimal forecasting, especially on predicting unusual passenger flow because it includes both local and global properties. SVM also works well with small datasets.
3. Multilayer perceptron using and Hybrid EMD-BPN strategy, can effectively reflect the characteristics of passenger flow and enhance forecasting accuracy, however mode mixing may appear. To overcome this, EEMD method can be applied to extract IMF components.
4. The LSTM model outperforms statistical and machine learning methods in terms of accuracy and stability, as it can effectively capture the nonlinear relationship and time dependency. The hybrid model shows great advantages when assessing peak passenger flow and is more adaptable to changes in bus passenger flow.
5. Traffic flow data can improve this type of prediction; however, this type of data can be used to gather (SCATS). Other types of traffic indicators may also be useful. Nevertheless, historical data models also give very approximate results. Temporal variables are also valuable.
6. Trip characteristics and demographic data can also be valuable inputs to predict public transports demand. Parking costs and availability or public transport accessibility index can also be considered important to the model.
7. Spatial-temporal correlation and specific scenario patterns influence positively bus traffic flow prediction model. Besides that, the model can be improved by considering point-of-interest data, social activities, and transportation networks.

The literature review indicates that there are many factors that influence population's behavior. Many different strategies have been applied to predict citizens fluency.

List of most common characteristics used to predict public bus flow:

- Variables that change in space and time:
    - Historical passenger flow (entries and exits)
    - Population movement (using different types of crowd data)
    - Traffic data (congestion level)
    - Weather
    - Other Events
    - Trip characteristics and demographic data
- Variables that change in time
    - Day of the week (weekend or not)
    - Rush hours
    - Day and Night
- Variables that change in space
    - Existence of near train stations
    - Existence of near metro stations
    - Number of distinct carriers that pass through that stop

2.7. MOBILITY REPORTS AND DASHBOARDS

To (Matheus, Janssen, & Maheshwari, 2018) the design of reports and dashboards should follow some basic principles:

- Collect accurate and precise data is crucial to have reliable information and make good decisions based on knowledge.
- Visualization should be customized according to the type of data and the main objective in study.
- Different views should be supported to improve understanding of the available information
- Clear presentation, charts should present clear and easy understandable information.
- Offer decision-making support, relating performance metrics with organizational measures to evaluate “What if” scenarios.
- Interaction support is one of the ways to offer users the possibility to investigate into data from different perspectives, giving them wider insights.
- Provide overview and details, giving them a wider and deeper perspective of the available information.
- Focus on creating public values, since they should be designed to create public values like engagement, transparency and accountability.
- Ensure real-time updates of data if possible, to increase confidence on possible insight and predictions made based on it.
- Ensure institutional support, guaranteeing trustworthy data.

Location based visuals are crucial, when analyzing a city performance, smart mobility analyses flows the same logic. Commonly this type of analyses makes use of location sensors (like GPS or other location capture technologies) that generate location data. Normally, data presented on a map, provides a better insight to use, since it makes information more comprehendible and perceivable.
According to Zheng, the location-based visuals involve three basic factors on the map:

- Type of location data.
- Visual forms.
- Data points representation.

The first one is directly associated with a business and its analysis. Geo locations displayed on a real-world map is one of the most used approaches. A second type of location data is local contextual locations, which do not directly rely on geo coordinates.
Real-world maps usually use a background layers, then points, paths and areas, displayed accurately or closely proximate to the background. Data points or paths forms and colors may influence the way that this project looks at data, they should follow the message that the presenter is trying to introduce and avoid any possible distraction from the visual objective.

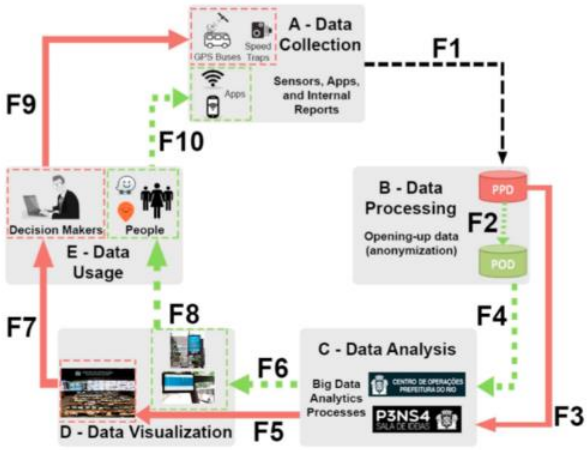

Figure 4 - Data Cycle for Dashboards (Matheus et al., 2018)

3. METHODOLOGY

To this project implementation, I will use the Design Science Research (DSR) method, since its main goal is the creation of an application environment, which culminates in finding a solution for a concrete problem. In this project, the application environment "will be the dashboards and will be developed based in the knowledge flows, process steps and outputs" (Sofia & Filipe, 2018).

This methodology has 4 main steps of implementation:

1. Awareness of Problem and Suggestion
    - This project will be critical to start changing mindset of event planning and management, based on multiple data sources to one single unified application environment that may address the problem as an all and optimize decision-making. A prototype suggestion will be generated, using metrics and perspectives based on acknowledgments retrieved on literature review.
2. Development
    - At this step, it will be defined the first framework of the possible solution regarding our problem. Some exploratory analyses will be made in order to identify improvement opportunities, taking in consideration the available resources.
3. Evaluation
    - At this stage, performance measures will be evaluated, to identify the application environment impact on the specified problem. It will also assess the possibility of addressing other related subjects, or even bring up to the table new problems solution as a collateral effect. Consequently, there is a possibility of having the need to apply some fine-tuning to the developed solution.
4. Conclusion
    - Exhibition of the project results and conclusions, taking in consideration what was previously defined, as the main objective of this research.

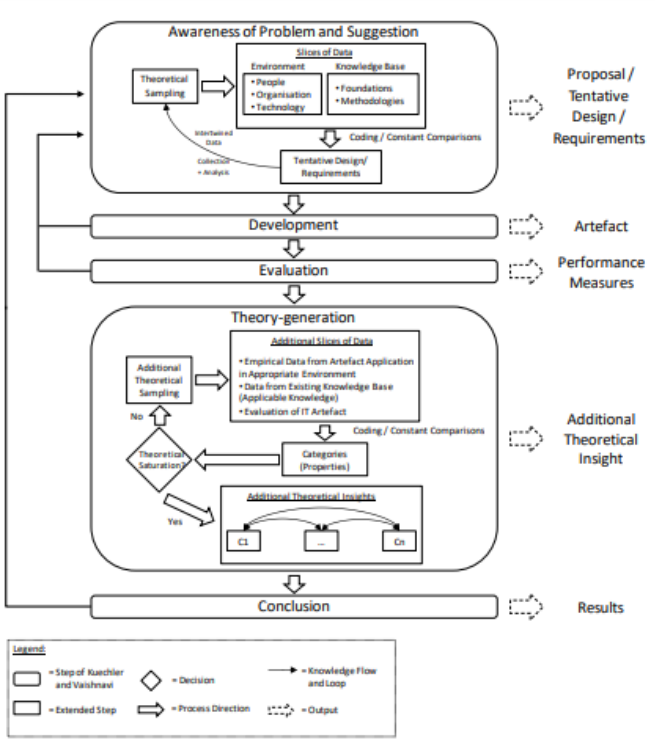

Figure 5 – DSR Process Model (Kuechler & Vaishnavi, 2011)

3.1. CONCEPTUAL MODEL PROPOSAL

In this segment the conceptual model proposal is presented. It is also called as the 'suggestion' step on DSR methodology. At this stage, the approach passes from creating a conceptual artifact versatile as possible in a way that it can receive data from any spatial-temporal situation.

It will start by giving an overview context of the case study available data and process flows, going through the creation of models, measures and metrics definitions that are based on literature review made previously.

At the end, the first conceptual model proposal is complete and ready to be materialized. Final step before entering "Development" section from DSR methodology.

This project model will be based on a structured database that will be used to compile all the data before starting data exploitation.

3.2. CONTEXT

Considering the application of the project on Lisbon Marathons crowd management, an overall overview of this specific type of event is given in this section.

This type of event occurs yearly and moves a vast amount of population increasing the need of smart decisions, regarding crowd mobility.

Marathons unlike most events, have an influence in a vast area since population that attend start at specific location but end at another very far way depending on the type of event (marathon, half marathon).

Besides the obvious massive population movement from a location to another, roadblocks are also a fact that should be taken in consideration. Roadblocks are normally necessary to ensure safety to marathon athletes. This is also having a huge positive impact over mobility, not just for runners but for all citizens.

In the next stage of this project a pipeline was created as a baseline. It starts with the creation of a Datawarehouse and its dataflow to extract, transform and load data into our data cube that will work as the data source of the final reports, by using direct query.

3.3. EXPLORATORY ANALYSIS PIPELINE

My strategy pipeline for the exploratory analyses will start by:

a) Creating a sample of events (marathons) on our area of interest in the temporal window mentioned above.

    a. Location

b) Use crowd's location from telecommunications cell towers data, to extract population concentration and movement in Lisbon center during the same period.

c) Use traffic congestion data from WAZE, that extracts jammed streets along the event day in Lisbon.

d) Making use of public transports data on the same space and time focusing on:

    a. CARRIS

        i. Enters

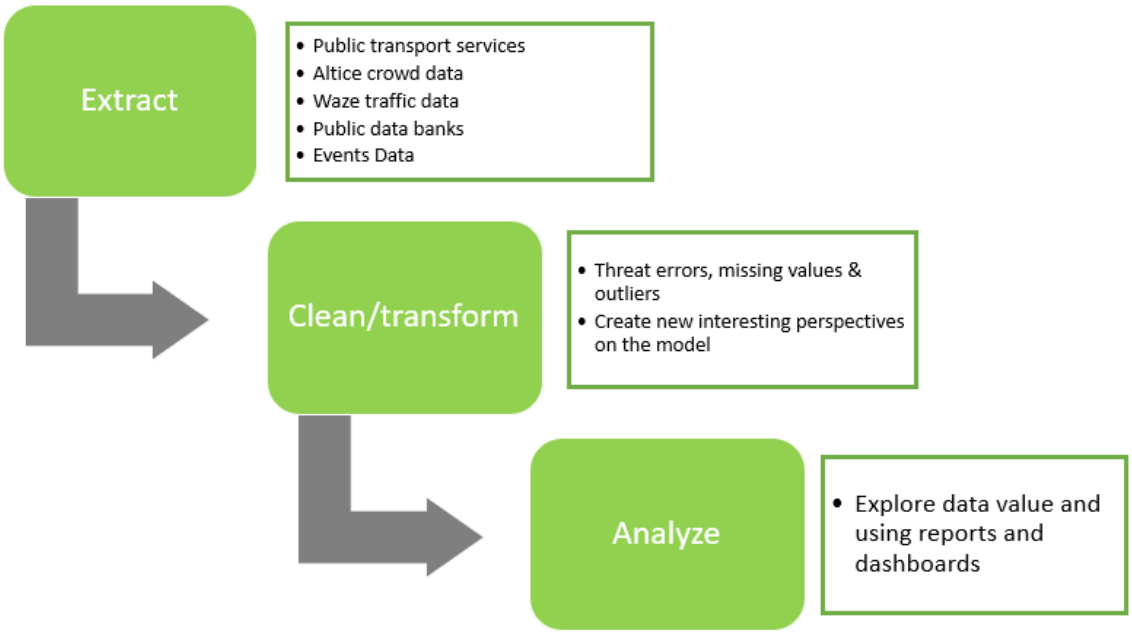

Figure 6 - Project Development Steps

Expected Outcomes (functional perspective):

- Based on a specific event:

  Define major tracks performed by population majority.

  - Improvement opportunities:

    Create new transport offers in that tracks or improve the existing ones.

  Points where people take their bus after leaving the event.

  - Improvement opportunities:

    Create new bus stops, or increase frequency before and after the event.

  Points of major traffic congestion before, during and after de event.

  - Improvement opportunities:

    Improve and publicize the existing transports offer in the same route, contributing to decrease traffic congestion.

Other Improvement opportunities:

If Lisbon have frequencies in a specific spatial-temporal environment, it may be able to measure the existing effort regarding demand and optimize this process.

Expected Outcomes solved (technical perspective):

- Find hot points during the event, for each type of transport.
- Find a pattern between the event and the type of transports, chosen by the event attendees.
  - Possible change of behavior by offering transportation in conjunction with the tickets.
    - Taking in consideration attendees features.
- KPI metrics that define critical moments of congestion in different types of transport and when there is a need of intervention.
- KPI metrics that define critical moments of congestion regarding traffic towards the event location.
- Create reports and dashboards that can be applied to other contexts with similar data.

3.4. CONCEPTUAL MODEL STRUCTURE

3.4.1. DATABASE MODEL

Data flow process will be organized in two central blocks:

- The first block will have basic processes, just to extract data from our sources (ex: csv files, json files) and apply some minor transformations.
- Second block will try to organize data in a clearer way to be able to extract knowledge and present data in a more readable and easier way.

This first part of the development will be made using SQL Server, SSIS, SSAS and Power BI software to extract, transform and clean data before start analyzing it. Threat missing values, errors and point out outliers.

The database model has source tables, destinations tables and views that directly feed the tabular model.

The tabular model will be the data source of our overall report, by using direct query Power BI option, that extracts data of our source in real time.

Database Tables:

- Source tables:
    - CROWD_LOCATION: Crowd timestamp data extracted from Altice csv files. Each csv represents a specific timestamp identified by the name of the file.
    - WAZE: Traffic data extracted from WASE excel files. File that reveals some insights regarding traffic congestion.
    - WAZE_CONTROL: WAZE data preparation table.
    - EVENTS_LOCATION: Events location and descriptive data extracted from json files. Events coordinates, date, and some related descriptive data.
    - METRO_SHAPES: Metro shapes location extracted from csv file.
    - METRO_STOPS: Metro stops location extracted from csv file. Stops coordinates.
    - METRO_ROUTES: Metro routes location extracted from csv file. METRO paths coordinates.
    - METRO_SHAPES_GROUP: METRO stops data preparation table.
    - METRO_STOPS_GROUP: METRO stops data preparation table.
    - CARRIS_SHAPES: CARRIS shapes location extracted from csv file.
    - CARRIS_STOPS: CARRIS stops location extracted from csv file. Stops coordinates.

    - CARRIS_ROUTES: CARRIS routes location extracted from csv file. CARRIS paths coordinates.
    - CARRIS_SHAPES_GROUP: CARRIS routes data preparation table.
    - CARRIS_STOPS_GROUP: CARRIS stops data preparation table.
    - CARRIS_MOVE: Entries on CARRIS stations.
    - ICON_TABLE: Icon related to data table.
    - EVENTS: Events data extracted from json files.
    - EVENTS_GROUP: Events data preparation table.
- Destination tables:
    - FACT_EVENTS_CROWD: Crowd location crossed with events.
    - FACT_EVENTS_PUB_TRANSP_SHAPES: Public transports shapes crossed with events.
    - FACT_EVENTS_PUB_TRANSP_STOPS: Public transports stops crossed with events.
    - FACT_EVENTS_WAZE: Waze traffic crossed with events.
    - ICON: Relevant Images.
    - DIM_DATE: Date dimension.
    - DIM_TIME: Time dimension.
    - PREDICT_CARRIS: CARRIS data aggregated by stop name.
    - PREDICT_CROWD: Crowd data aggregated by stop name.
    - BASE_PREDICTION_FINAL: CARRIS data aggregated by stop name and 30 minutes intervals.

Database Views:

- VDIM_DATE:
    - Date dimension.
- VDIM_TIME:
    - Time dimension.
- VDIM_WAZE:
    - Waze dimension.
- VDIM_EVENTS:
    - Events dimension.
- VDIM_LOGOS:
    - Logos dimension.
- VDIM_PUBLIC_TRANSP_SHAPES:

    - Public transports shapes dimension.
- VDIM_PUBLIC_TRANSP_STOPS:
    - Public transport stops dimension.
- VFACT_EVENTS_CROWD:
    - Events joined with crowd facts.
- VFACT_EVENTS_WAZE:
    - Events joined with traffic facts.
- VFACT_EVENTS_PUBLIC_TRANSPORTS_SHAPES:
    - Events joined with public transport shapes facts.
- VFACT_EVENTS_PUBLIC_TRANSPORTS_STOPS:
    - Events joined with public transport stops facts.
- VFACT_PREDICT_CARRIS:
    - CARRIS facts aggregated by stop name (area).
- VFACT_PREDICT_CROWD:
    - Crowd facts aggregated by stop name (area).
- VIEW_BASE_PREDICTION:
    - CARRIS and Crowd facts aggregated by stop name (area) and 30 minutes intervals (time).

3.4.2. DATABASE TABULAR MODEL

The database tabular model was built on top of the data warehouse, to have an independent layer between the data ingestion process and the reports, a method that enhances data control and consistency.

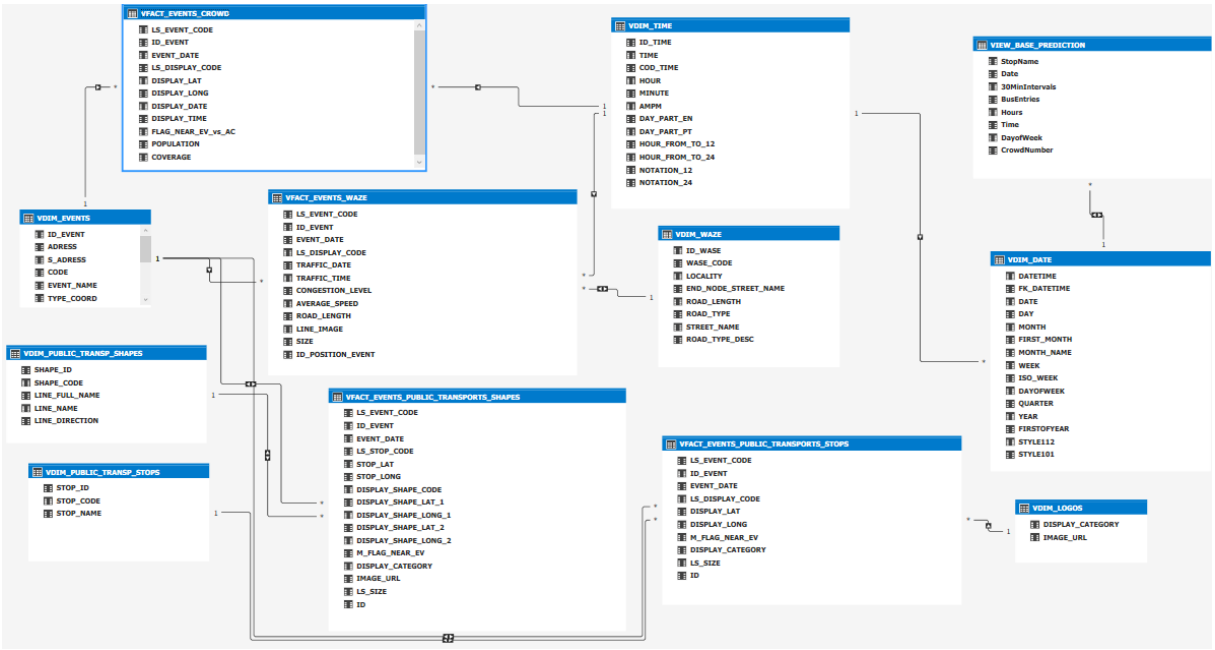

Figure 7 – Smart Lisbon Tabular Model

3.4.3. MEASURES

A list of measures was created based on the available data and the literature review most import notions.

- CARRIS_ENTRIES: Number of entries in a CARRIS bus
    - Type: Integer
- Occupancy: Peoples concentration on the event day (crowd hits).
    - Type: Integer
- Occupancy Day Before: Peoples concentration on the day before the event day (crowd hits).
    - Type: Integer
- Occupancy Day After: Peoples concentration on the day after the event day (crowd hits).
    - Type: Integer
- Public Transport Lines: Count of public transport Lines.
    - Type: Integer
- Public Transport Stop: Count of public transport stops.
    - Type: Integer
- Carris Stops: Count of Carris stops.
    - Type: Integer
- Metro Stops: Count of Metro stops.
    - Type: Integer
- Num Roads High Traffic: Number of roads with high level of traffic near the event.
    - Type: Integer
- Num Roads Medium Traffic: Number of roads with medium level of traffic near the event.
    - Type: Integer
- Num Roads Low Traffic: Number of roads with low level of traffic near the event.
    - Type: Integer
- Length High Traffic: Total length covered by high traffic roads.
    - Type: Integer

- Length Medium Traffic: Total length covered by medium traffic roads.
    - Type: Integer
- Length Low Traffic: Total length covered by low traffic roads.
    - Type: Integer
- Total Num Near Roads: Sum of all roads near the event.
    - Type: Integer
- Num Roads High Traffic (Actual -1): Number of roads with high level of traffic near the event (1 day before the event).
    - Type: Integer
- Num Roads High Traffic (Actual +1): Number of roads with high level of traffic near the event (1 day after the event).
    - Type: Integer
- % High Traffic: Number of high levels of traffic roads (hits) near the event per total number of existing roads near event.
    - Type: Percentage/Float
- % Medium Traffic: Number of medium levels of traffic roads (hits) near the event per total number of existing roads near event.
    - Type: Percentage/Float
- % Low Traffic: Number of low levels of traffic roads (hits) near the event per total number of existing roads near event.
    - Type: Percentage/Float
- Route number Count:
    - Count of CARRIS Entries without any previous aggregation.
- Crowd Actual Day:
    - Sum of Crowd on the Actual Day
- Crowd Day Before:
    - Sum of Crowd on the Previous Day
- % Population TD vs YD
    - Comparison of Crowd between the actual day and the day before.
- BusEntries TD
    - Sum of CARRIS Entries this day

- BusEntries TW
    - Sum of CARRIS Entries this week
- Bus Entries TD vs PW
    - Comparison of CARRIS Entries Between this day and the previous week.

4. EXPLORATORY ANALYSIS RESULTS

4.1. DASHBOARDS & REPORTS

Considering all literature review and the available data, a Power BI report called Smart Lisbon Mobility was developed, to join Crowd, Traffic, and Public Transport data with Events data. The report was built on top of the previous mentioned cube, using a direct query to force always having the most updated data available. All report pages were made, to present data in a simple and self-explanatory way.

The “Smart Lisbon Mobility during big Events” report was built to give an overview of some key aspects about the population’s movement across the city, during Lisbon marathons.

The report follows the same structure across all pages, suing the same structure, fonts and colors taking in consideration its main objective.

Map visuals are also vital contemplating the fact that this project objective is included in smart cities concept.

All pages have filters which grant users the possibility to explore different perspective.

4.1.1. Crowd Management Section

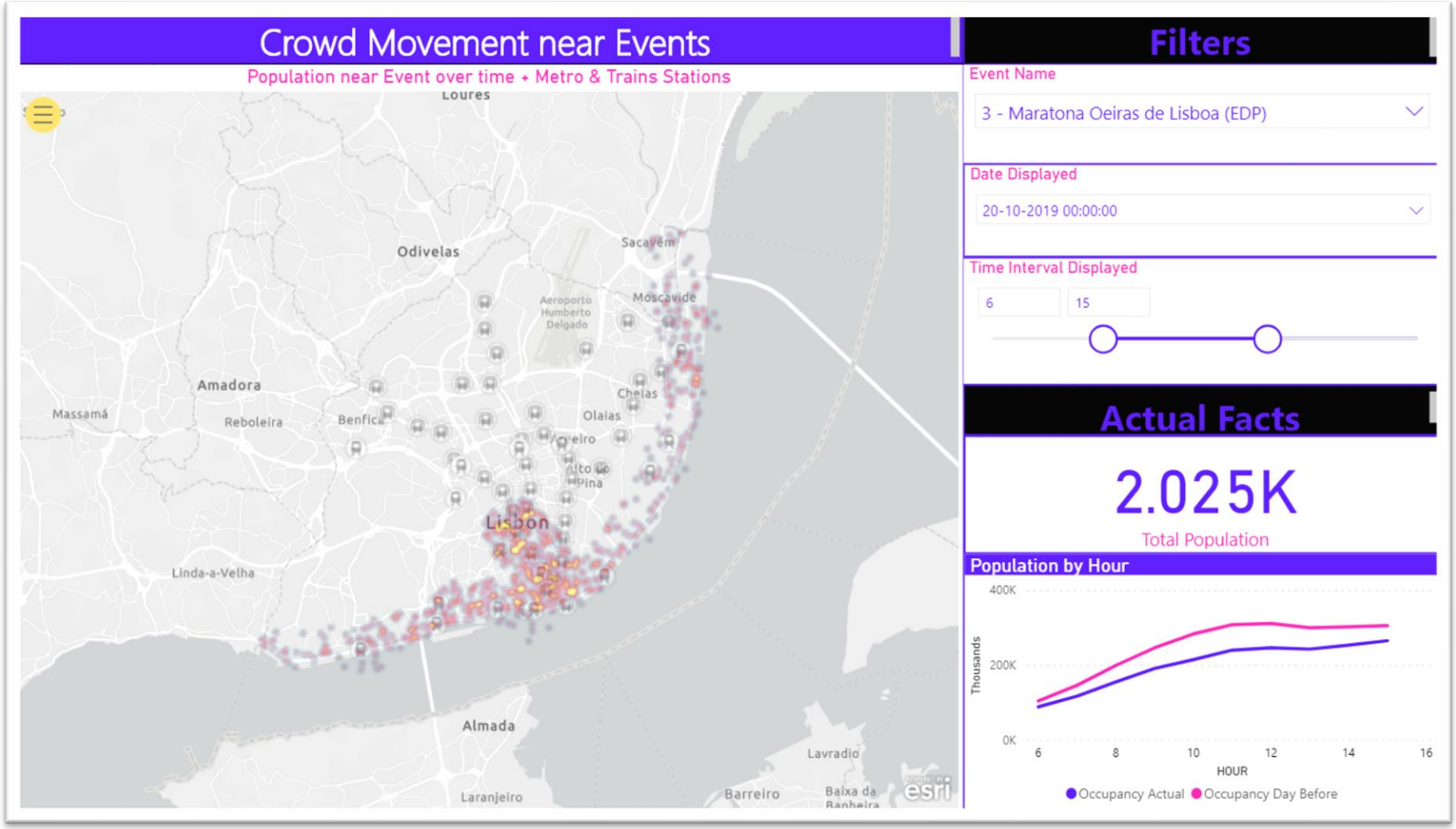

Figure 8 - Crowd Management I

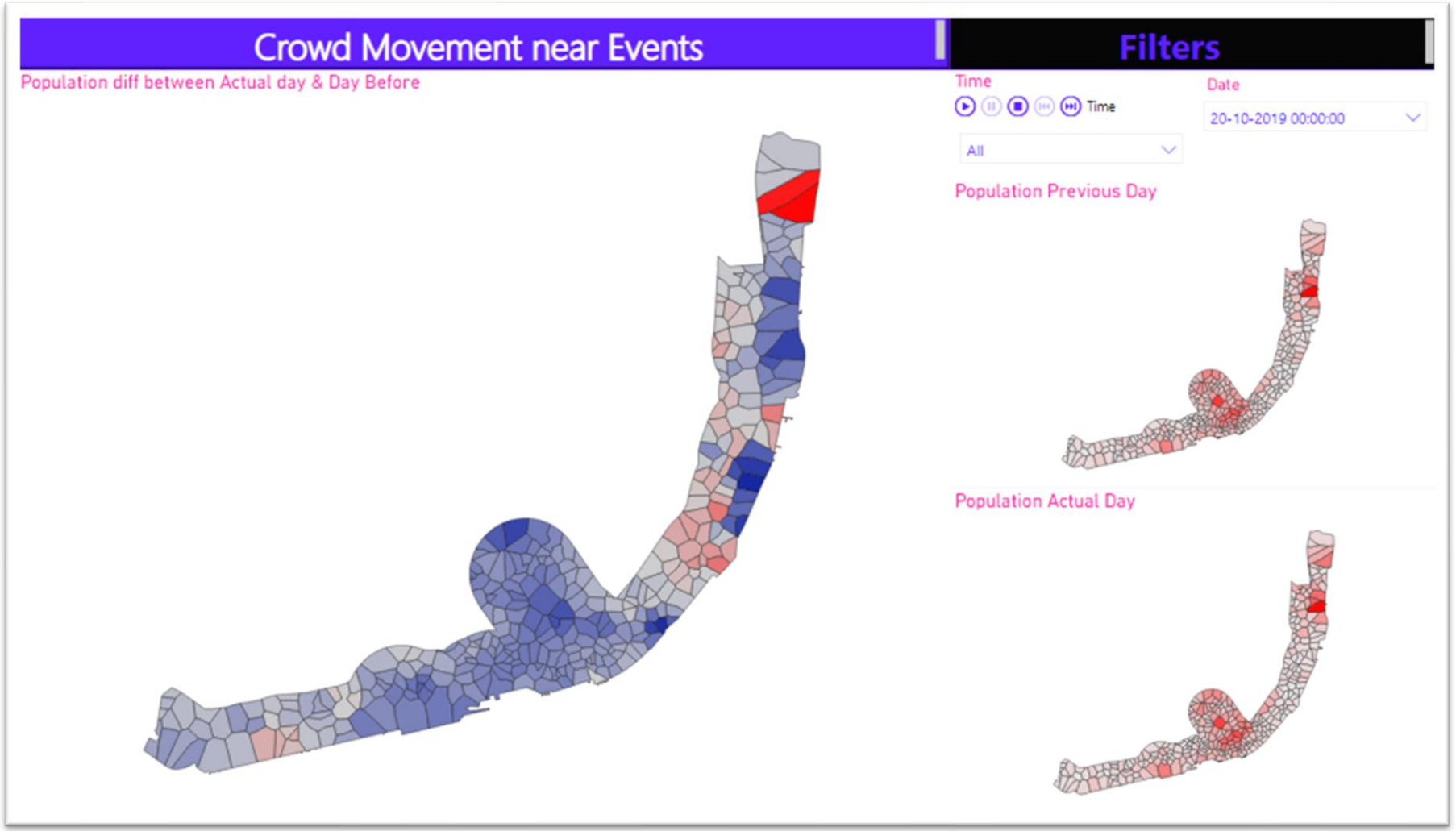

Figure 9 - Crowd Management II

This 1st section aims crowd management analyses. They use population movement tracking, using “Altice” cell towers during to extrapolate mobility patterns across the city before, during and after the event.

On the 1st page, a heat map layer is displayed using ArcGIS maps (Light Grey Canvas). A Metro and Train station location layer is also displayed on the map. Some slicers were built to focus our analyses on a specific event, day, and time. On the bottom right it is also displayed a Card with the total population represented on that page and a line chart comparing the same population volume by hour with the previous day.

Blue < Yellow < Red → Color used to represent population concentration on the heat map.

On the 2nd page, there are 3 shaped maps aggregating data by stop names. A date and time slicers were added. A Play Axis was also added to give users the opportunity to monitor the evolution throughout timestamps. On the left side of the report there is the primary shape map representing population % diff between the actual day and the previous day. On the right, the additional 2 shaped maps represent population concrete figures. The top one represents the previous day and the bottom the selected day.

Blue < Grey < Red → Color used to represent population % diff between actual and previous day on the main heat map.

Grey < Red → Color used to represent population actual figures on both secondary shape maps (actual figures).

4.1.2. PUBLIC TRANSPORTS MANAGEMENT SECTION

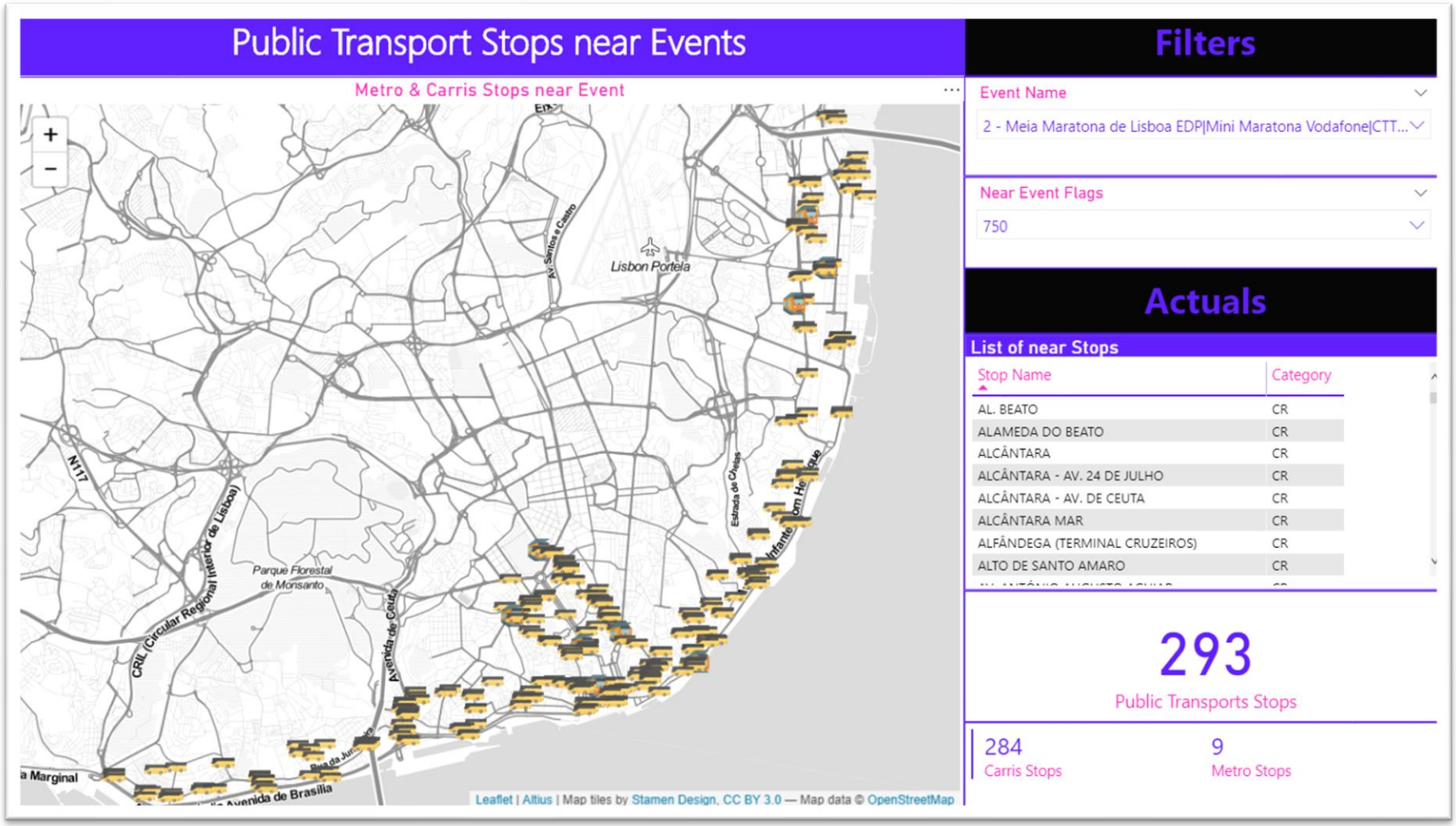

Figure 10 – Public Transports Stops

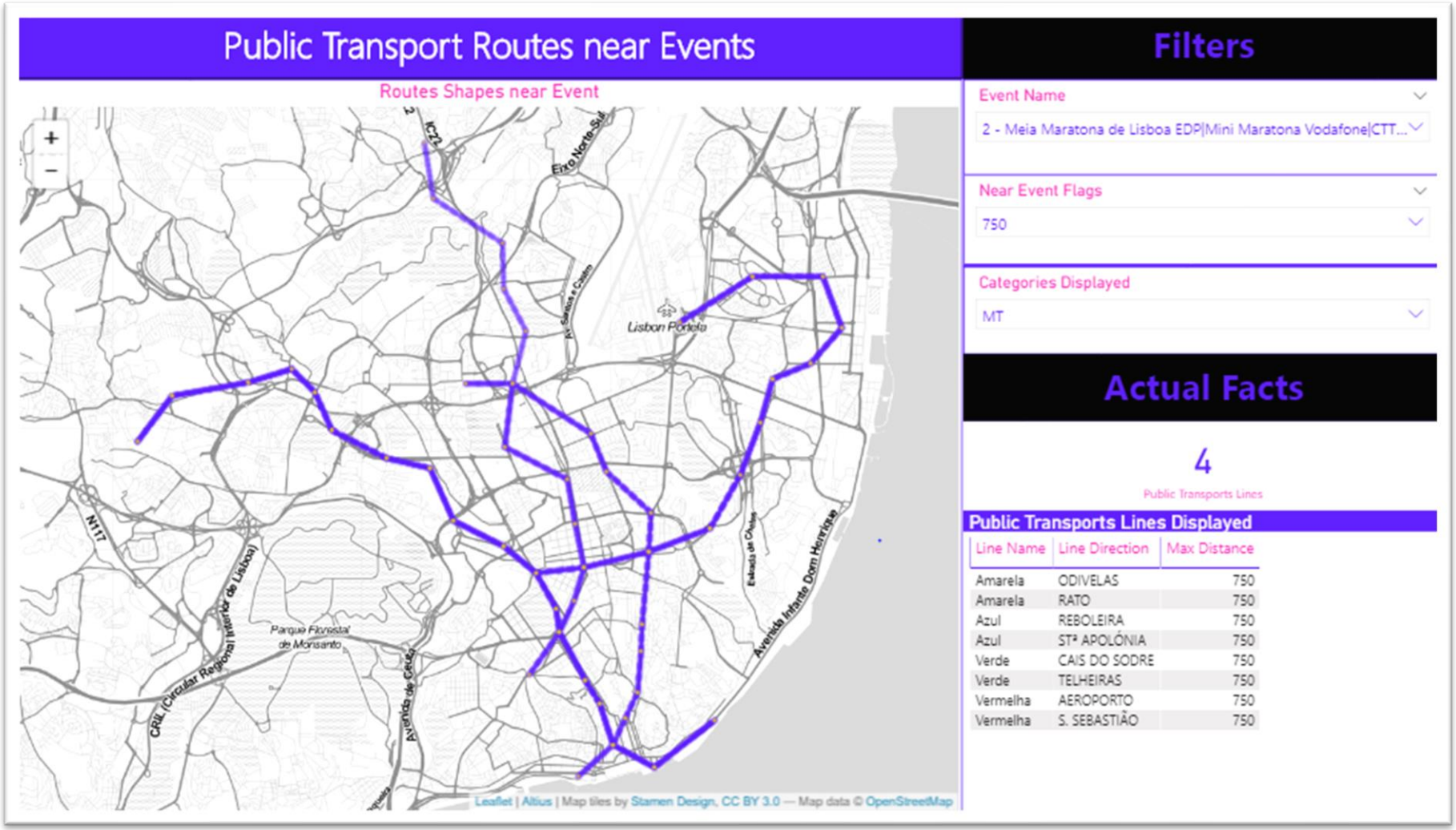

Figure 11 – Public Transports Shapes

This 2nd section focuses on public transports management, for this project case, it concentrates on CARRIS and METRO analyses. It uses stops and routes location to extrapolate the reach of each point of the city.

On the 1st page of this section, an icon map for bus and metro stations is displayed over the area of study (marathon area). This map uses Stamen – Tonerlite background. Some slicers are available to select the event, the definition of nearness, where it was selected at 750 meters, corresponding to the threshold found on our literature review. Some complementary infor is presented on the bottom right. There is a table cataloging all stops names and its category, as well as a card with the total number of stops displayed and a multi-row card with the total number stops aggregated by category (CARRIS and METRO).

Yellow bus represents CARRIS and the orange Metro represents METRO. Icons images can be modified on icons database table.

On the 2nd page, there is a second icon map for bus and metro route shapes (lines) that have at least one station near the study area of influence (marathon area). This map also uses Stamen – TonerLite background. Some slicers are once again available to select the event, the definition of nearness and the transport category. Some complementary info is presented on the bottom right. There is a card with the total number of lines displayed and a table cataloging all the line names, line direction and the threshold that it is included.

4.1.3. TRAFFIC MANAGEMENT SECTION

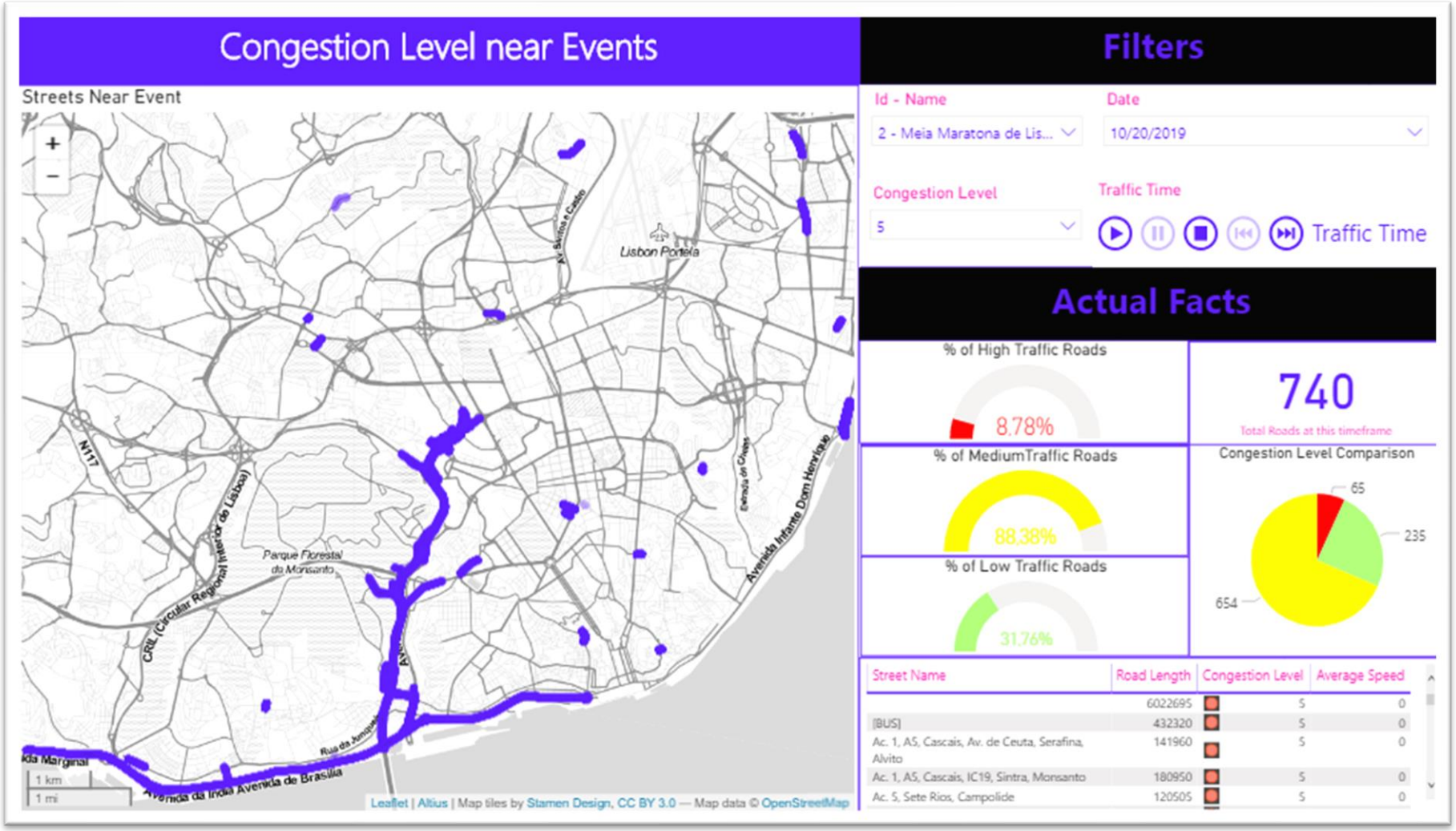

Figure 12 – Traffic Management I

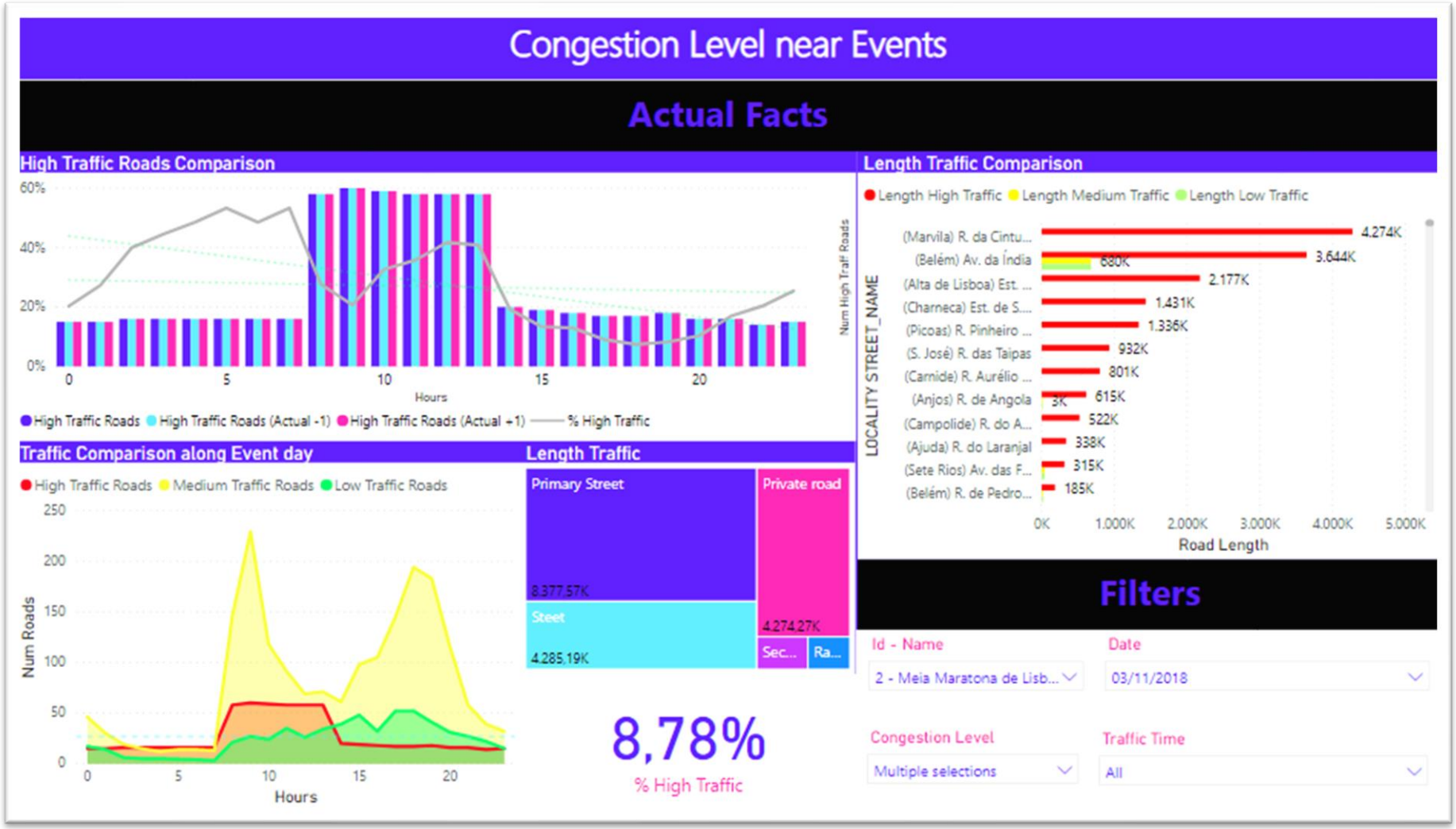

Figure 13 - Traffic Management II

This 3rd section centers on traffic management near the event. It uses WAZE data to control traffic before, during and after the event retrieving insights.

On the 1st page of this section, an icon map for jammed roads is displayed. This map uses Stamen – TonerLite background. Some slicers are available to select the event, date, the congestion level (where 1 represents low traffic and 5 road that the road literally is blocked). There is also a play axis available giving users the possibly to monitor changes across the time. On the bottom right some complementary facts are presented. There are 3-gauge charts representing total number of roads with high, medium, and low traffic. A card with the total number of roads displayed within the timeframe in case and a pie chart to compare these totals. There is also a table listing all this street names, their length, congestion level and average speed.

The red represents high traffic roads, the yellow medium traffic roads and the green low traffic roads.

On the 2nd page, there are available some more visuals to give a more in-depth comprehension over traffic congestion near the event. This page has a line and clustered column chart with the number of high traffic roads on the actual day, the previous day, and the day after. On the bottom left there is an area chart comparing the total number of high traffic roads, medium traffic roads and low traffic roads by hour. On the top right corner there is a clustered bar chart representing length traffic comparison (high, medium, and low) by locality, street name or road type. On the middle there is a tree map representing length traffic by road type and a percentage of high traffic roads card. After all, on the bottom right, a list of slicers is available to select a specific event, congestion level, date and time.

Red represents high traffic roads, Yellow medium traffic roads and green low traffic roads.

All the other colors are used according to the report design and are only used to represent a category, meaning that their color is not meaningful by itself.

4.1.4. CARRIS MANAGEMENT SECTION

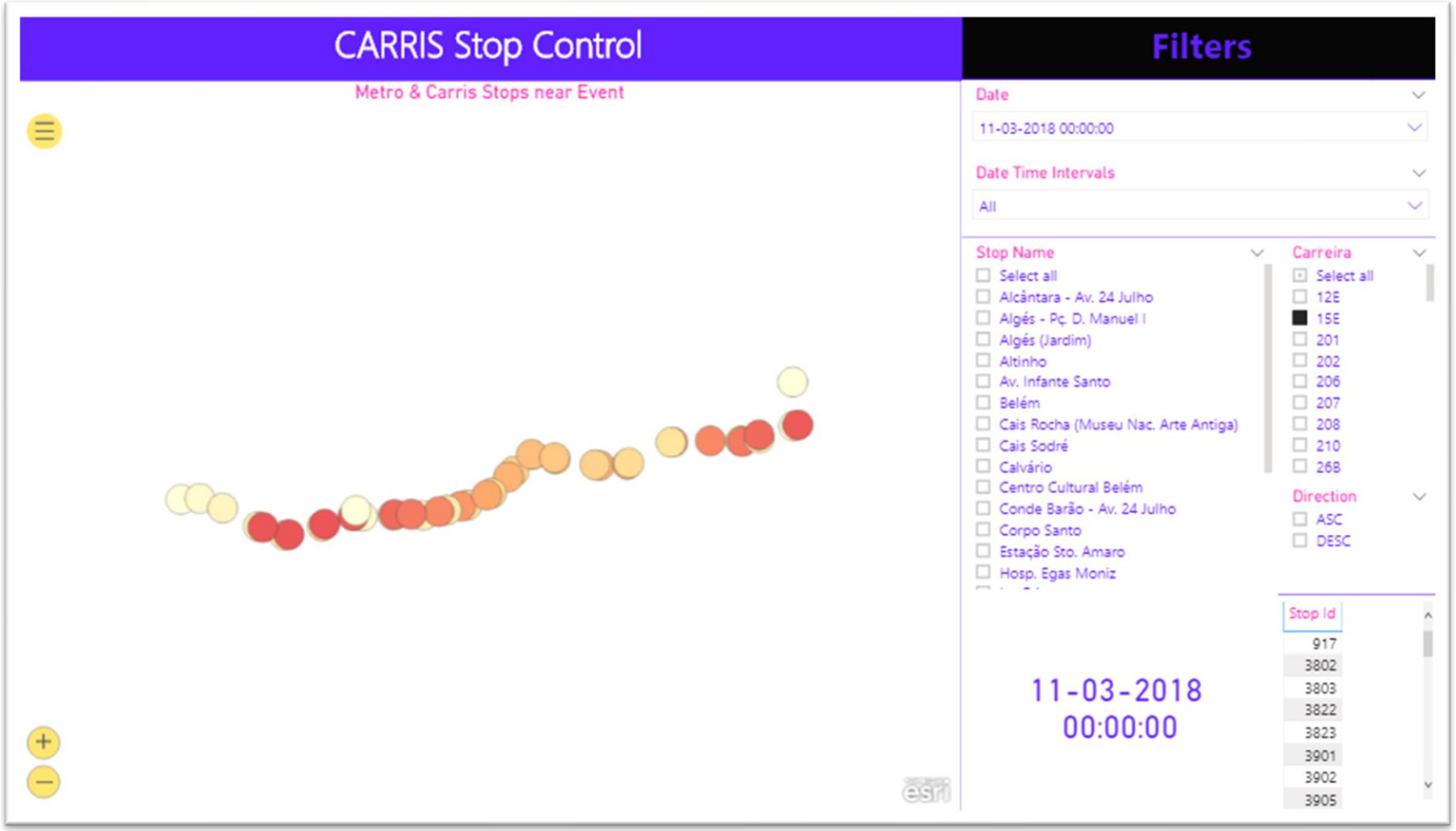

Figure 14 – CARRIS Management I

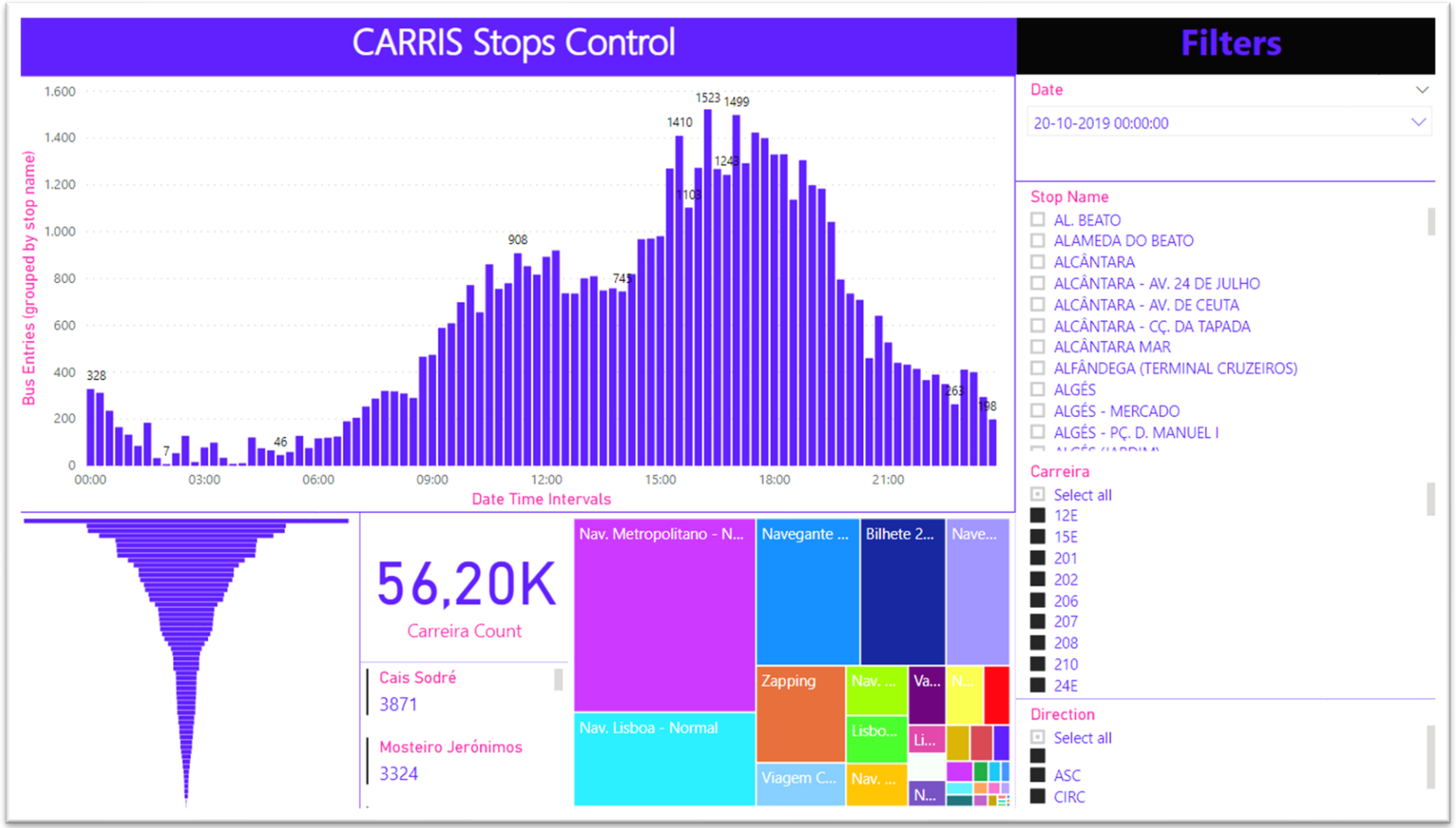

Figure 15 – CARRIS Management II

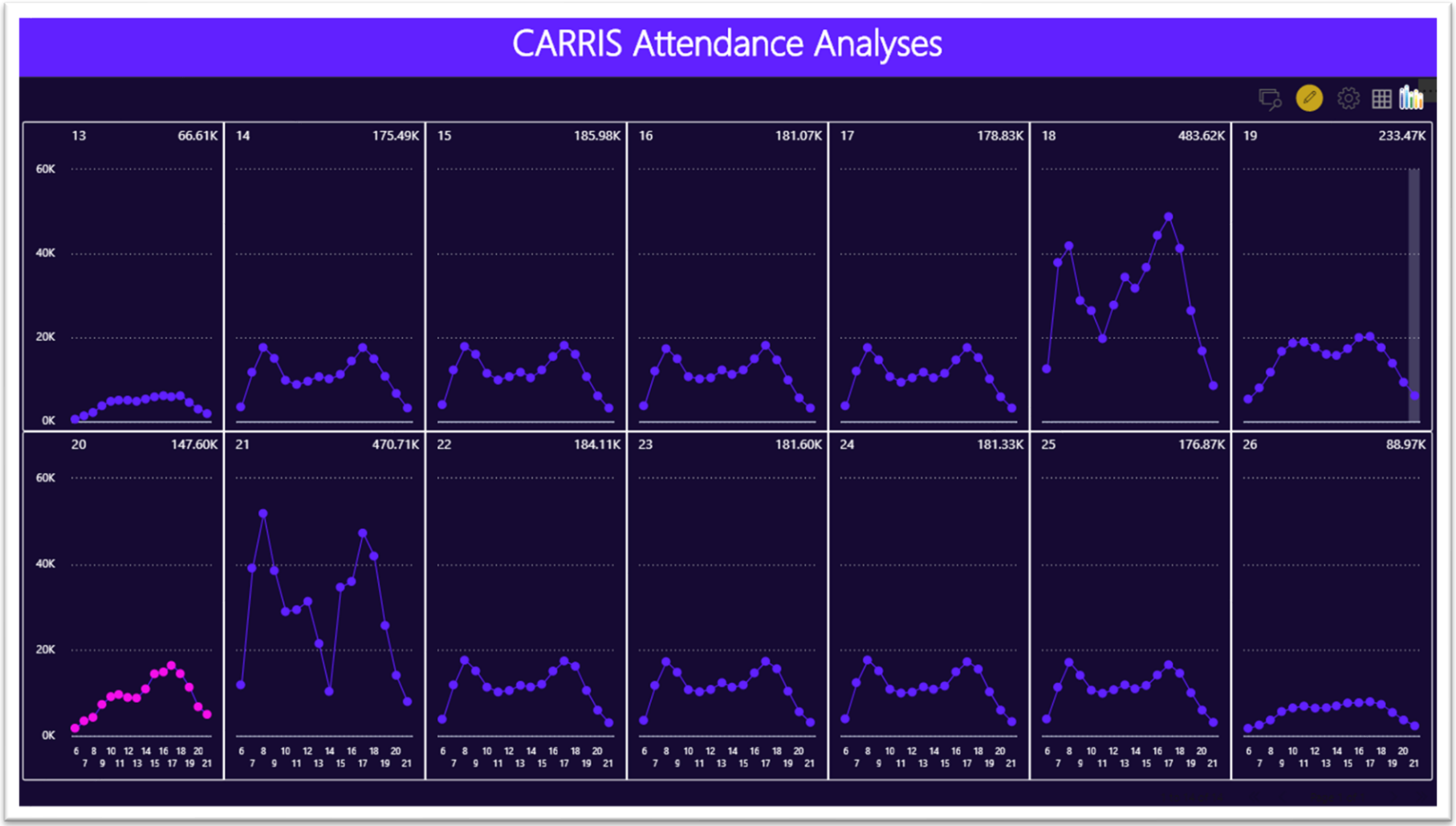

Figure 16- CARRIS Management III

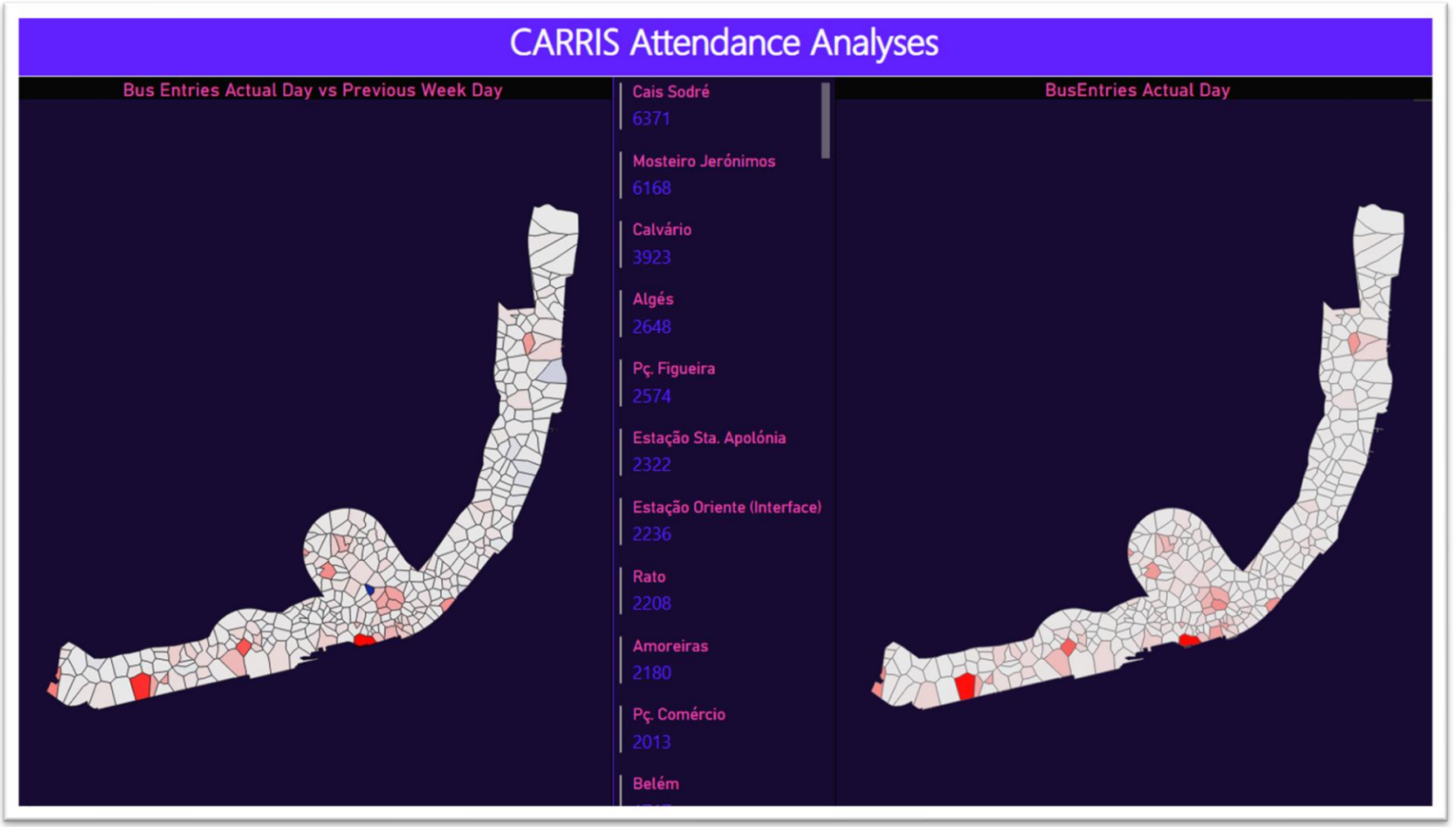

Figure 17 - CARRIS Management IV

This 4th section focuses on CARRIS management, during marathons. It uses CARRIS data to find population normal behaviors.

On the 1st page, a colour map layer is displayed, using ArcGIS maps (Light Grey Canvas). On the left side, a map chart is visible representing bus entries for each stop id (bubble) where its colour indicates the total of entries. On the top right corner there are a few slicers to aim our analyses on a specific date, time interval, stop name, route number and direction. On the bottom right there is a datetime card and a table list of all displayed id stops as complementary information to the report.

Beige < Yellow < Orange < Red → Colors used to represent population concentration on the color map.

On the 2nd page, a more in-depth analysis is performed. On the top left there is a column chart representing CARRIS bus entries, aggregated by stop name and datetime intervals for a predefined day. On the bottom left there is a funnel representing bus entries for each route number. On the bottom middle there is a card represent the total number of entries on the day in study, a multi-row card representing the total number of entries on that same by “Stop Name”, and a tree map representing bus entries for each type of ticket. On the right side, a list of slicers is displayed to analyse by date, stop name, route number, and direction.

On the 3rd page, an Advanced Trellis – xViz chart is available, offering the possibility to find visual bus entries shapes across days and hours of that period in study. The pink line represents the marathon day.

On the 4th page, 2 shape maps and a multi-row card is displayed to give a clearer view of the total number of bus entries on the marathon day. On the left, the shape map compares bus entries on the marathon day with the same period of the previous week. On the right side we have the total bus entries on the marathon day (actual day). On the middle the multi-row card describes those same actual figures by stop name (ordered by actuals).

4.1.5. CARRIS vs CROWD Section

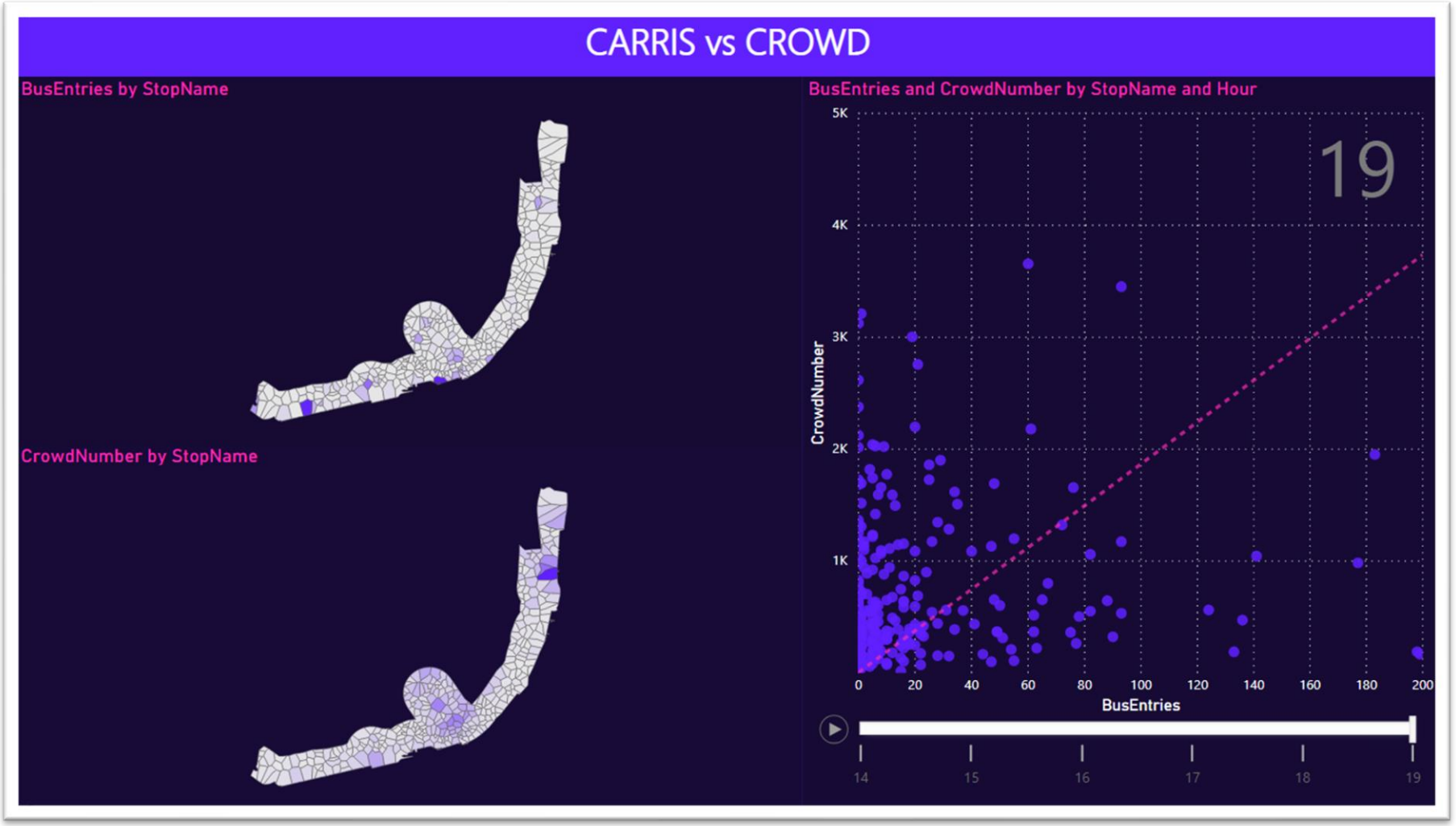

Figure 18 - CARRIS vs. CROWD

This section has one page giving an overview user an overview of the relationship between Bus Entries and Crowd Number. It joins CARRIS and CROWD data aggregating those by stop name. On the top left side of the page there is shape map representing Bus Entries on the day of the marathon for each stop name. Below, there is Crowd concentration with the same aggregation to compare both reports. Both reports use the same scale and colours. On the right side there is a scatter chart comparing crowd and bus entries by hour. A ratio line is also displayed to give a better perspective of what line should data point follow in order to be considered highly correlated (positive).

4.2. EXPLORATORY ANALYSIS CONCLUSIONS

To better understand marathons events and its impact on Lisbon mobility, the marathon that took place on 20/10/2019 (Sunday), sponsored by EDP, that started at 8AM in Cascais and was planned to finish at 12PM was the leading event selected.

After having our first event defined, an exploratory analysis has been made using Smart Lisbon Mobility report (report mentioned on the previous section).

On the 1st and second pages (crowd management section) it was concluded that:

- There was a crowd increase between the beginning of the marathon (8AM) and the end (midday).
    - Nevertheless, after checking the other days and comparing them, this increase start looking like a pattern.
    - However, we notice that there is a decreasing of population on the event day comparing with the other days.
        - Probably because many streets are closed for athletes' passage.
- Analyzing *Praça do Comércio* zone (marathon goal line) in depth, it is visible that there was an increase of population, comparing with the previous day and the day after the event.
- Along the route there was a decrease of population, possibly because of streets closure contemplating athletes' safety.
    - On the first and second page (crowd management section) it was concluded that:

On the 3rd and 4th pages (public transports management section) it was inferred: that:

- Along the zone in study (marathon path with a max 7,5 minutes using Euclidean distance) there are 293 stops (METRO and CARRIS).
    - 284 from CARRIS
    - 9 from METRO
- However only 8 stops are is near *Praça do Comércio (Baixa-Chiado*) end of the line.
    - 7 from CARRIS
        - 714 – *Outurela*
        - 714 – *Praça da Figueira*
        - 728 – *Portela*
        - 728 – *Restelo*
        - 736 – *Cais do Sodré*
        - 736 – *Odivelas*
        - 759 – *Estação Oriente*
        - 759 – *Sul e Sueste*
        - 15E – *Alges*
        - 15E – *Praça da Figueira*
        - 711 – *Praça do Comercio*
        - 711 – *Alto Damaia*
    - 1 from METRO

        - *Baixa-chiado*
            - Blue & Yellow lines
            - 2 directions each

On the 3rd and 4th pages (traffic management section) it was extrapolated that:

- The percentage of road blocked increase on the day of the event due to road closure for athletes passage as mentioned before.
    - We can clearly see the map the path used since the streets are only blocked during marathon timeline.
    - This influences all road traffic, including public transport routes which used to pass through those streets that were blocked.
        - *Belém (Av. da India)* is the longest road more affected.
        - Most of the jammed streets are primary streets.
- The peak hours are between 7AM and 14PM.
- Comparing with the previous day and the day after, the peak hours are the same.
    - Suggested by the increased number of both high and medium traffic roads.

On the 5th, 6th, 7th and 8th pages (CARRIS management section) it was extrapolated that:

- On both marathons, peak hour is between 14PM and 19PM
- There is a clear change of demand between weekdays and weekends
    - This indicates that people use mainly this transport to go to work or school.
    - There is also a small difference between Saturdays and Sundays
    - Saturday tend to have more demand
- While trying to extrapolate data from 2018 marathon, it was concluded that:
    - The 3 most requested route numbers on the event day are 728, 15E and 28E, by that order.
    - CARRIS as registered 51 810 Bus Entries
        - From those, the majority used Navegante Urbano Normal has ticket card (12,800)
    - *Cais Sodré, Mosteiro Jerónimos and Praça Figueira* are the most requested, by that order on 11/03/2018 (Event Day)
    - There is a Bus Entries peak between 14h and 19h (hours after the marathon) comparing with the previous week.
        - On this specific time interval, stops most crowded were *Mosteiro Jerónimos, Cais Sodré and Praça Figueira*, by that order.
            - All these locations have registered more than a thousand bus entries.
            - *Mosteiro Jerónimos* was the event end of the line.
- While applying the same analyses to 2019 marathon it was inferred that:
    - The 3 most requested ROUTE NUMBERS on the event day are 15E, 760 and 28E, by that order.
    - CARRIS as registered 56 200 Bus Entries
        - From those, the majority used Navegante Urbano Normal has ticket card (15,812)

- *Cais Sodré, Mosteiro Jerónimos and Calvário* are the most requested, by that order on 20/10/2019 (Event Day)
- There is a Bus Entries peak between 14h and 19h (hours after the marathon) comparing with the previous week.
    - On this specific time interval, the stop most crowded was *Cais Sodré*
        - This stop was near the event end line (*Praça Comércio).*
        - *Praça Comércio* stop registered 2 thousand entries.

4.2.1. CARRIS VS CROWD DATA

Pearson correlation was retrieved between Bus Entries and Crowd count, in order to clarify if both variables are correlated. The metric told us that these variables have a positive correlation of 21%. It cannot be assumed that they are highly correlated. Nevertheless, the opposite is also not true. To have a better understanding, a few more test have been made during this project using charts that were meant to draw their relationship. If these variables were correlated enough, this could point out that one could help predicting the other using machine learning algorithms.

5. PREDICTION MODEL DEVELOPMENT

In this section, the prediction model for CARRIS demand plan is put in place. The model will focus on marathon days in order to anticipate demand during this out of the box occasions. The model should be able to predict demand before, during and after the event. To do this, a machine learning model was built over Bus Entries by stop name and hour.

5.1. DATA PREPROCESSING

Spatial aggregation:

- First, data was aggregated in space, by stop name, reducing the number of stops areas to 343, focusing on the marathon's paths.

Temporal aggregation:

- Data was aggregated to have date/time intervals of 1 hour. Bus Entries were summed according to those intervals.

The base dataset had 29 features and 109 200 records based on this project exploratory analysis, literature review and available data.

The features used were the following:

- *HourIntervals*
    - Date/time value representing hour intervals.
        - YYYY/MM/DD hh:mm:ss
        - Ex: 2018/03/08 14:00:00 represents the sum of all bus entries between 2018/03/08 14:00:00 and 2018/03/08 15:00:00.
- *StopName*
    - String representing a group of stops aggregated according to their location.
        - Ex: Cais Sodré or Mosteiro dos Jerónimos.
- *BusEntries*
    - Integer number representing the sum of bus entries for that stop name and hour.
    - This will be the dependent variable.
- *Date*
    - Date format (part of *HourIntervals* features)
        - YYYY/MM/DD.
- *Hours*
    - Integer representing hours of the day
- *Time*
    - Time format (part of *HourIntervals* features)
        - hh:mm:ss
- *DayofWeek*
    - Categorical integer representing day of week.
        - Starts with 1 representing Sunday and ends with 7 representing Saturday.
- *FlagEvent*

    - Binary variable
        - 1 when it is the event day (marathon) and 0 if not.
- *FlagMorning*
    - Binary variable
        - 1 if we are in the morning (between 6h and 11h59) and 0 if not.
- *FlagAfternoon*
    - Binary variable
        - 1 if we are in the afternoon (between 12h and 16h59) and 0 if not.
- *FlagEvening*
    - Binary variable
        - 1 if we are in the evening (between 17h and 22h00) and 0 if not.
- *FlagNight*
    - Binary variable
        - 1 if we are in the night (between 22h and 05h59) and 0 if not.
- *FlagWeekend*
    - Binary variable
        - 1 if we are in the weekend (Saturday or Sunday) and 0 if not.
- *FlagRushHour*
    - Binary variable
        - 1 if we are in the rush hour (between 8h and 9h59 or between 16h and 18h59) and 0 if not.
- *FlagEventPeak*
    - Binary variable
        - 1 if we are in the event peak hour (between 14h and 18h59) and 0 if not.
- *CountGiraStops*
    - Integer representing the number of GIRA stations.
        - Value that changes only in space (stop name).
- *FlagMarathonPath*
    - Binary variable
        - 1 if we are in the marathon track (stop name) and 0 if not.
- *CountJams*
    - Integer representing the count of jams (roads blocked)
        - Value that changes in space and time
- *MeanSpeed*
    - Float representing the average speed
        - Value that changes in space and time
- *CountTrainStops*
    - Integer representing the number train stations
        - Value that changes only in space
- *CountMetroStops*
    - Integer representing the number metro stations
        - Value that changes only in space
- *Temperature*
    - Integer representing the temperature

        - Value that changes in space and time
- *Humidity*
    - Integer representing the temperature
        - Value that changes in space and time
- *Precipitation*
    - Integer representing the temperature
        - Value that changes in space and time
- *Sun*
    - Integer representing the temperature
        - Value that changes in space and time
- *WindSpeed*
    - Integer representing the temperature
        - Value that changes in space and time
- *WindCardinal*
    - Categorical integer representing cardinal point.
        - Starts with 1 representing North and ends with 8 representing North-West.
        - Value that changes in space and time.
- *Route numbersCount*
    - Integer representing the number of distinct CARRIS carriers by StopName area
    - Changes only in space
- *StopsCount*
    - Integer representing the number of distinct CARRIS stop stations by StopName area
    - Changes only in space

To reach the best base dataset, all records that had *FlagNight* equal to 1 (between 22h and 5h59) where excluded from the dataset ending with 77 350 records. This decision was made according to the literature review since night data would only bias our results. *FlagNight* feature was also excluded from the dataset finishing with 28 features.

5.2. MODEL IMPLEMENTATION

5.2.1. IMPORT SECTION

The model starts by importing the base dataset csv file mentioned previously into a data frame.

5.2.2. CHECK MULTICOLLINEARITY

In this section, multicollinearity is checked to understand if from all the available independent features, there are some that can be excluded, to avoid redundancy (Table 7).

For each iteration, the feature with the highest VIF value was removed from the dataset, until we only had features with a value lower than 5 (VIF conservative threshold). At this point the model had 17 variables.

5.2.3. FEATURE SELECTION

In this step, feature selection procedure is applied using RFECV. This method is applied using ridge linear model, as estimator. Cross validation is applied with 5 slices.

The model concluded that 9 features were enough to explain the event.

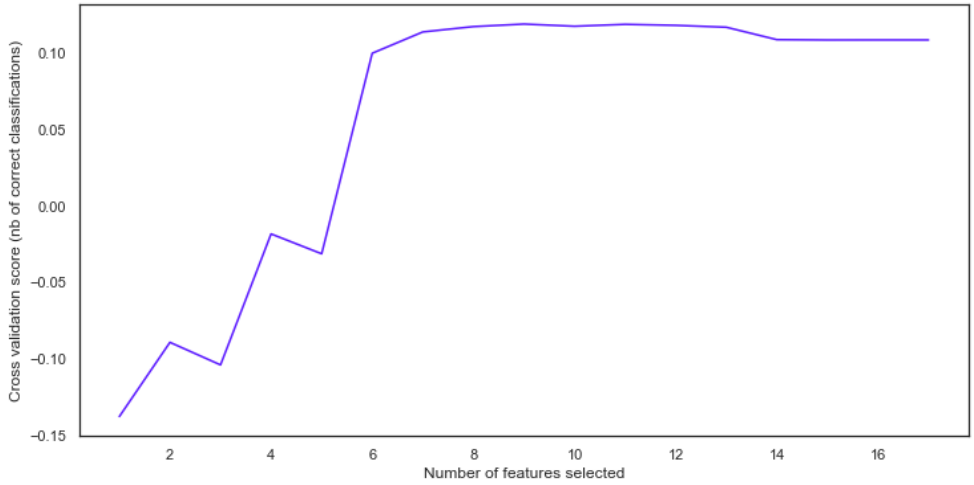

Figure 19 – Optimal number of features (RFECV)

The model also defined a ranking for all the 17 features in study selecting only 9 highlighted in the table below:

DayofWeek	Selected False	Rank: 6
FlagEvent	Selected False	Rank: 2
FlagMorning	Selected True	Rank: 1
FlagAfternoon	Selected True	Rank: 1
FlagWeekend	Selected True	Rank: 1
FlagRushHour	Selected True	Rank: 1
FlagEventPeak	Selected False	Rank: 3
CountGiraStops	Selected True	Rank: 1
MarathonPathFlag	Selected True	Rank: 1
CountJams	Selected False	Rank: 7
MeanSpeed	Selected False	Rank: 4
CountTrainStops	Selected True	Rank: 1
CountMetroStops	Selected True	Rank: 1
Precipitation	Selected False	Rank: 8
Sun	Selected False	Rank: 9
WindSpeed	Selected False	Rank: 5
Route numbersCount	Selected True	Rank: 1

Table 6 – Feature Importance Rankings (RFECV)

At the end of this section only 9 features were nominated by RFECV for the final prediction model created further.

5.2.4. Train and Test split

Before applying any algorithm, a train test split is applied ending with 75% of records (58 012) on train side and 25% on test (19 338). This split was made with a stratification parameter on both *FlagEvent* and *StopName* variables. This way, it was guaranteed that we had the exact same proportion of event records over all stop names in both train and test dataset.

5.2.5. ALGORITHM SELECTION

The next step in the pipeline, was the algorithm selection. To do that 4 algorithms were compared using the square error (R^2) as a scorer. To do this cross validation with 5 splits was applied on the train dataset.

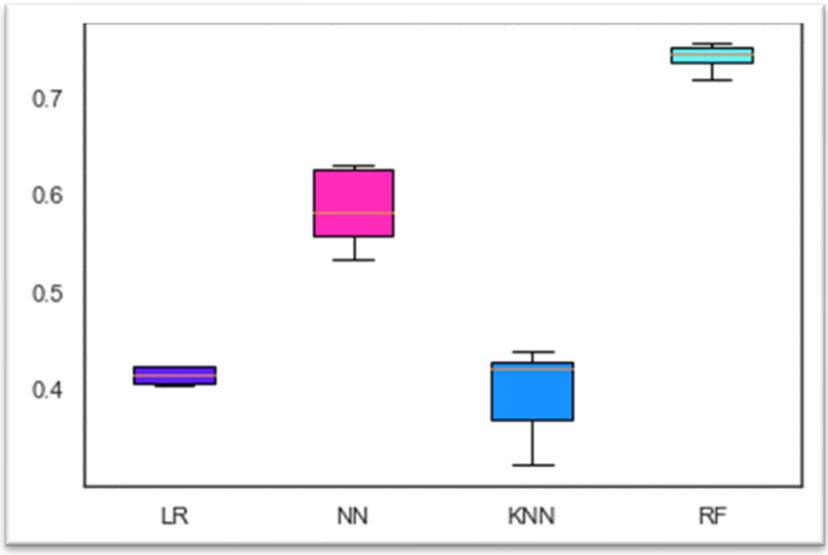

Figure 20 – Algorithm Comparison (LR: Linear Regression, NN: MLP Regressor, KNN: K Neighbours Regressor, RF: Random Forest Regressor)

Random Forest algorithm was defined the best model with R^2 of 0,747279 and a standard deviation of -0,012918.

Algorithms	Linear Regression	Neural Network	K-Neighbors Regressor	Random Forest
R^2	0.412786	-0.361705	0.405601	0,747279
Standard Deviation	0.013179	1.310844	0.044399	-0,012918

5.2.6. TRAIN MODEL

In this step of the pipeline, the model is trained using GridSearchCV to find the best parameters. Grid Search defined that the best parameters to apply on Random Forest algorithm are the following:

- Number of estimators = 1000
- Max Features = Log2
- Max Depth = 9

The scorer used by the model was RMSE. The model was trained with cross validation with 5 splits.

5.2.7. TEST MODEL

To finalize the pipeline, the model was tested using test dataset and the best model defined on the previous section (using the parameters recommended). Prediction results were exported to a csv file

that would be ingested further by the process that loads data into the database and ends it on Power BI. The measures used to test the model were based on the literature review.

5.3. MODEL RESULTS

On Smart Lisbon Mobility during big Events report, a final page was added to include Prediction Results insights.

Explained Variance (R^2)	MSLE	MAE	MAPE	MSE	RMSE
75,60%	1,7841	8,3383	181,7298	912,9388	30,2149

Table 7 – Prediction Model results using the test dataset

The model was able to predict 75,60% Bus Entries (dependent variable). The average real error is 30 bus entries. Mean average percentage error is 182% each may reveal that the errors do not follow a normal distribution. To validate that, a distribution error chart was produced:

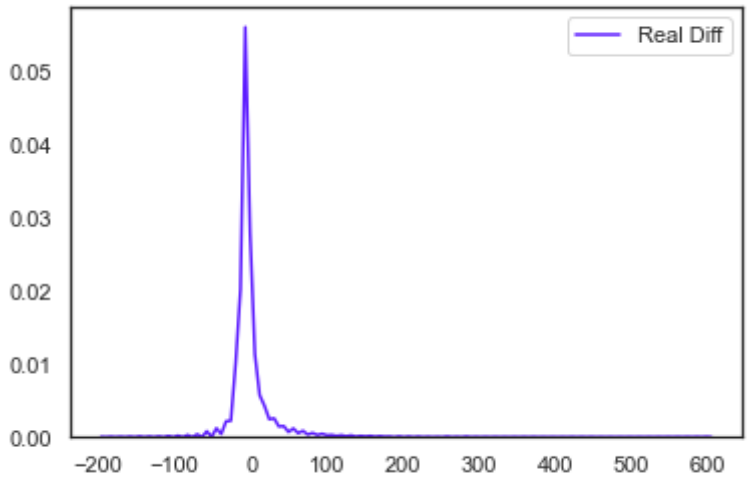

Figure 21 – Distribution Error Chart (test dataset)

The chart indicates that the model follows a normal distribution with a small deviation to the negative side. Although the tail of the distribution seems to be too wide, which explains the high percentage error of the model. There are a few records where the model completely failed increasing the MAPE.

When analysing the Prediction Results Section focusing on the event day (11-03-2020) it was concluded that:

- The highest errors, using real difference, were on *Mosteiro Jerónimos*, *Santa Justa – Rua Ouro* and *Cais Sodré* (all with differences above 350).
- The highest errors, using percentage difference, were on *Santa Justa – Rua Ouro* and *Glória – Restauradores*.

- Comparing Sundays with weekdays, the model underestimates on Sundays and overestimates on all the other days. However, the error is always less than 20% in average per day. On the event day the model underestimated with 15% of error in average.

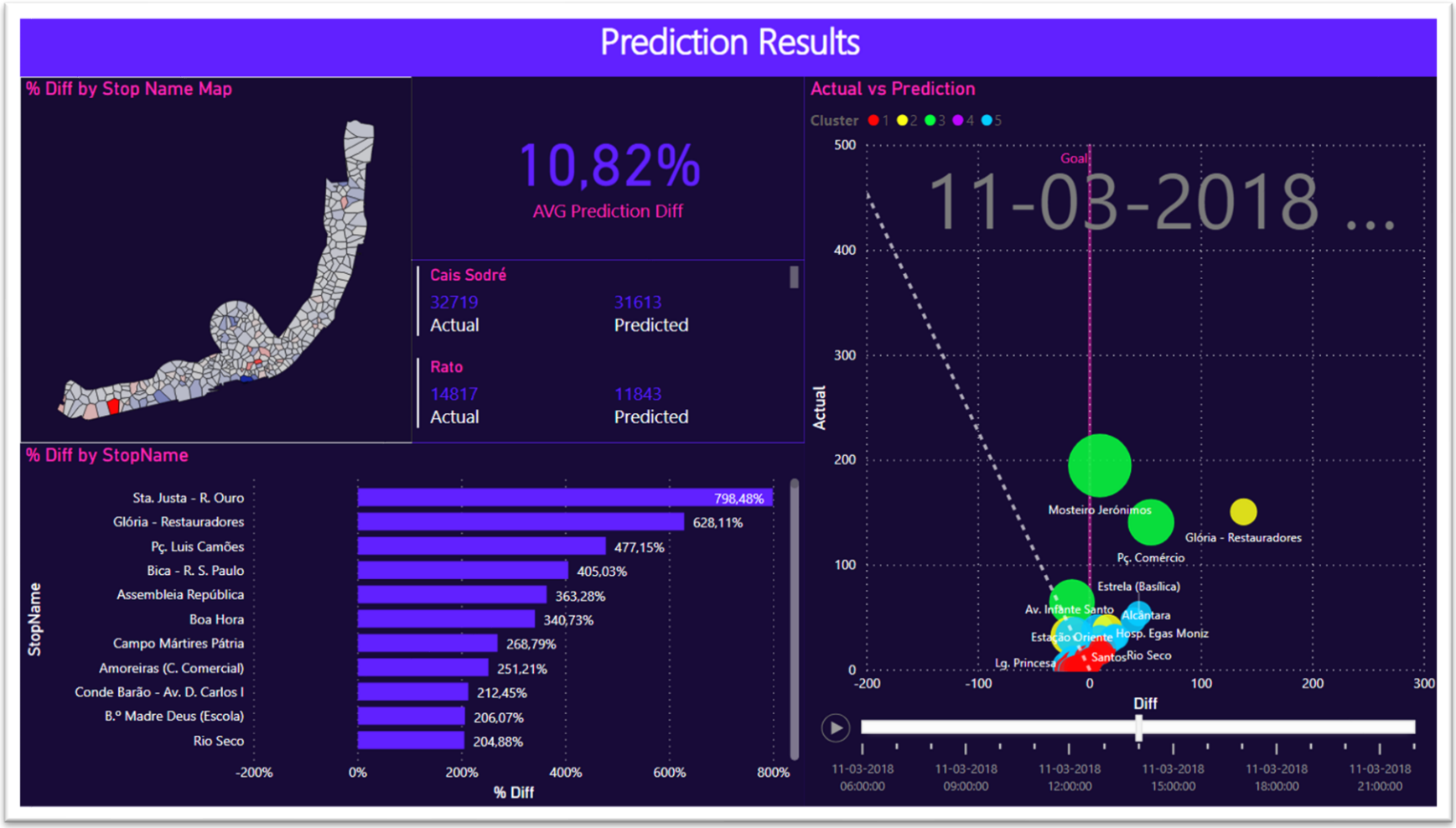

Figure 22 – Prediction Model Results

6. CONCLUSIONS

It was concluded that a machine learning model can be created to predict CARRIS attendance during marathons for each stop name, by using multiple sources of data and a process to extract, threat and load it all into a database model that culminates with reporting.

This type of model can only be built over an exhaustive exploratory analysis of the entire environment in study. Monitoring the city behavior for a long period over multiple perspectives using multiples sources of data enlisted as one, is best way to reach solid base for any prediction model. All this can be integrated as one and run in an automated way as seen with the solution made for this project. With the integration of multiple sources of data into a single database it is also possible to examine other smart city aspects from different perspectives, such as traffic, crowd, and event management. The ETL model was built to work in a fast and automated way, although it still has plenty of room for improvement. A data cube was built and can be used to deliver a self-service platform across Power BI, to give users the possibility to create their own reports, according to their individual needs. By building a platform that besides giving predictions also offers past insights based on facts, the model offers the opportunity to evaluate decision making plans effectiveness. Integrating all kind of key influencers of smart city mobility can increase the entire model quality. Having better results on CARRIS would also increase other public transports entities service quality. Is more than known that public transports have a direct relationship between each other for multiple reasons, that give this project confidence enough to corroborate that smart mobility concept, that is only achievable if all entities work as one pointing to the same goal. That goal should be the improvement of all citizens life quality by giving an optimized public transport system. When speaking of big events, such as marathons, public transports enter in an uncomfortable zone, since behaviors can be totally different from what is called a usual day. Considering this, this project built a prediction model that would focus on those specific cases. This project was built not only to monitor when and where are the stops and routes most requested, but also predict future cases, especially when they are in odd days, such as marathons. The CARRIS prediction model was successfully built and approved by members of the entity itself (CARRIS) that considered this path the future of the public transports system, agreeing with the entire project main message. It is possible to plan better, having predictions as a core influencer during decision making, instead of intuition. CARRIS may reallocate efforts where it is most needed reducing costs and increasing life quality of all their clients, by granting a more sophisticated offer, even during atypical happenings such as a marathon. Besides giving citizens a better transport offer, CARRIS would also benefit from it if they reduce costs, by optimizing their service while reallocating their efforts when and where they are most needed. The model already gives good insights, predicting 75% of bus entries, even though, the model still has margin to improve. Considering everything said until now, the main question is already answered, the better quality of life regarding smart mobility is achievable by building a prediction model above a knowledgeable exploratory analysis platform that incorporates the key factors of citizens behavior. A model should be implemented not only to one entity but to all, since they all serve the same purpose and the same clients, in this case Lisbon citizens and possible foreigners. A model that should be simple and intuitive, to help all sorts of managers take better decisions (event managers, public transport managers, crowd managers). It is important to mention that the entire structure presented is only the base, the tip of the spear is always the decision-making entities. The main difference is that until this moment decision was mostly based on sense and intuition and from now on it can be based on knowledge retrieved from data.

7. LIMITATIONS AND RECOMMENDATIONS

The model still has a vast margin to be improved. One feature that may be important is crowd data from Altice. To this project there was no available crowd data during the marathon in study. Other features that may be useful are METRO, CP, GIRA, and any other types of public transports attendance since they are all related. A considerable number of people use multiple types of transport to get to their destination. Using events data, can also be a path even though it may be hard to access those, specially when speaking of unplanned events.

Training the model with one event and testing it with a second would also improve the model. The results reports would give more confidence, in the sense that better insights could be retrieved when analyzing the error across all hours, weekdays and stop names, instead of having to apply a train test split over a single event.

Applying the same process over all stop stations instead of area (group of stations) would also be an improvement. This model only concentrates on stop areas (*StopName*) for performance reasons. This model ran on only one machine, making the unfeasible the idea of running the model for every single stop station. A model could be implemented using a *Distributed File System* methodology such as *HDFS*. This way the model would run much faster providing insights with a higher frequency.

The ETL process can also be optimized, to reduce run time and to include the prediction script made in python. By having all this process integrated, a schedule can be created to have the most updated insights, ready for CARRIS decision making, through Power BI reporting, without any human interaction.

A time-series prediction model was also built, but it was concluded that it was unachievable since the available data had a 1-year gap between events. This type of model requires a continuous dataset to give good quality insights. This is a path that can also be useful in this type of predictions. Results can be compared to machine learning models to understand if both models reach the same conclusion reinforcing the model level of confidence.

8. BIBLIOGRAPHY/REFERENCES

Addison, J. D., & Heydecker, B. (2015). Modelling Traffic with Variable Speed Limits. *Transportation Research*, (June).

AML. (2020). Metropolitan Area of Lisbon Portal. Retrieved 29 June 2019, from AML website: https://bit.ly/2IXP9nO

Anna, J. E., Abbott, L., & Geddie, M. W. (2001). Event and Venue Management : Minimizing Liability Through Effective Crowd Management Techniques EVENT AND VENUE MAN : MINIMIZING LIABILITY THR EFFECTIVE CR. *Event Management*, *6*, 259–270.

Asmael, N., & Waheed, M. (2018). Demand estimation of bus as a public transport based on gravity model. *MATEC Web of Conferences*, *162*, 1–5. https://doi.org/10.1051/matecconf/201816201038

Barbosa, M. de A. (2018). 70 mil pessoas e outros números do Web Summit. Retrieved 29 June 2019, from Sapo website: https://bit.ly/2GxhHTX

Benevolo, C., Dameri, R. P., & Auria, B. D. (2016). Smart Mobility in Smart City Action Taxonomy , ICT Intensity and Public Bene fi ts. In *Lecture Notes in Information Systems and Organisation*. https://doi.org/10.1007/978-3-319-23784-8

Connors, B. E. (2007). *Planning And Managing Security For Major Special Events : Guidelines for Law Enforcement*. Washington, D.C.

Daraio, C., Diana, M., Di Costa, F., Leporelli, C., Matteucci, G., & Nastasi, A. (2016). Efficiency and effectiveness in the urban public transport sector : a critical review with directions for future research. *European Journal of Operational Researc*, *248*(March), 1–20. https://doi.org/10.1016/j.ejor.2015.05.059

David, C. M. (2017). Lisboa é a cidade ibérica com mais trânsito. Mas está entre as melhores da Europa em mobilidade. Retrieved 20 June 2019, from Público website: https://bit.ly/2XfmILd

Ewen, J. (2019, October 25). How Big Data Is Changing The Event Planning & The Events Industry. Retrieved 9 December 2019, from Tamoco website: https://bit.ly/2sbgbCm

Han, Y., Wang, C., Ren, Y., Wang, S., Zheng, H., & Chen, G. (2019). Short-term prediction of bus passenger flow based on a hybrid optimized LSTM network. *ISPRS International Journal of Geo-Information*, *8*(9). https://doi.org/10.3390/ijgi8090366

Heldens, S., Litvak, N., Steen, M. Van, & Senior, I. (2018). Scalable Detection of Crowd Motion Patterns. In *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*.

Jain, S., Ronald, N., Thompson, R., & Winter, S. (2017). Predicting susceptibility to use demand responsive transport using demographic and trip characteristics of the population. *Travel Behaviour and Society*, *6*(January), 44–56. https://doi.org/10.1016/j.tbs.2016.06.001

Khozium, M. O., Abuarafah, A. G., & AbdRabou, E. (2012). A Proposed Computer-Based System Architecture for Crowd Management of Pilgrims using Thermography. *Life Science*, *9*(2), 277–282.

Klauser, F. (2012). Spatialities of security and surveillance : Managing spaces , separations and circulations at sport mega events. *Geoforum*, *49*, 289–298. https://doi.org/10.1016/j.geoforum.2012.11.011

Kuechler, B., & Vaishnavi, V. (2011). Promoting Relevance in IS Research : An Informing System for Design Science Research Rigor and Relevance : A Recurring Dilemma. *Informing Science: The International Journal of an Emerging Transdiscipline Volume*, *14*.

Larsson, A., & Ranudd, E. (2019). *The Analysis of Pedestrian Movement and Behaviour of Different Crowds during Stadium Egress*. Lund University.

Leung, I. X. Y., Chan, S., Hui, P., & Li, P. (2011). *Intra-City Urban Network and Traffic Flow Analysis from GPS Mobility Trace*. Cambridge.

Li, Y., Wang, X., Sun, S., Ma, X., & Lu, G. (2017). Forecasting short-term subway passenger flow under special events scenarios using multiscale radial basis function networks. *Transportation Research Part C: Emerging Technologies*, *77*(April), 306–328. https://doi.org/10.1016/j.trc.2017.02.005

Liu, P., Zhang, Y., Kong, D., & Yin, B. (2019). Improved spatio-temporal residual networks for bus traffic flow prediction. *Applied Sciences (Switzerland)*, *9*(4). https://doi.org/10.3390/app9040615

Lopes, S. P. (2018). *Lisbon Smart Cities : Perception and Reality*. Nova Information Management School.

Lusa, & Público. (2019). Transporte de passageiros aumenta no segundo trimestre. Retrieved 9 December 2019, from Público website: https://bit.ly/35cohcM

Matheus, R., Janssen, M., & Maheshwari, D. (2018). Data science empowering the public : Data-driven dashboards for transparent and accountable decision-making in smart cities. *Government Information Quarterly*, (November 2016), 101284. https://doi.org/10.1016/j.giq.2018.01.006

Mazloumi, E., Currie, G., Rose, G., & Sarvi, M. (2009). Using SCATS data to predict bus travel time. *32nd Australasian Transport Research Forum, ATRF 2009*, 1–14.

Mynatt, E., & Tullio, J. (2001). *Inferring Calendar Event Attendance*.

Neto, M. de C., & Cartaxo, T. de M. (2020). Algorithmic Cities: A Dystopic or Utopic Future? In *Intelligent Systems, Control and Automation: Science and Engineering* (pp. 59–73). Retrieved from https://link.springer.com/chapter/10.1007/978-3-030-56926-6_6

Pan, G., Qi, G., Zhang, W., Li, S., & Wu, Z. (2013). *Trace Analysis and Mining for Smart Cities : Issues , Methods , and Applications*. (June), 120–126.

Pires da Costa, A. H. (2008). *Manual do Planeamento de Acessibilidades e Transportes: Transportes Públicos*. Porto.

Rao, A. M., & Rao, K. R. (2012). *MEASURING URBAN TRAFFIC CONGESTION – A REVIEW*. *2*(December). https://doi.org/10.7708/ijtte.2012.2(4).01

Redação. (2019). Novos passes em Lisboa: guia para saber tudo o que muda. Retrieved 9 December 2019, from idealista website: https://bit.ly/2E33kVu

Sampaio, B. R., Neto, O. L., & Sampaio, Y. (2005). *Efficiency Analysis of Public Transport Systems: Lessons for Institutional Planning*.

Schaffers, H., Komninos, N., Pallot, M., & Trousse, B. (2011). Smart Cities and the Future Internet: Towards Cooperation Frameworks for Open Innovation. In *Future Internet Assembly*.

Heidelberg: Springer.

Sofia, C., & Filipe, P. (2018). *DESIGN OF A SUPPLY CHAIN MANAGEMENT Information sharing to mitigate the bullwhip effect*. Nova Information Management School.

Wei, Y., & Chen, M. C. (2012). Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks. *Transportation Research Part C: Emerging Technologies*, *21*(1), 148–162. https://doi.org/10.1016/j.trc.2011.06.009

Xue, R., Sun, D. J., & Chen, S. (2015). Short-term bus passenger demand prediction based on time series model and interactive multiple model approach. *Discrete Dynamics in Nature and Society*, *2015*(i). https://doi.org/10.1155/2015/682390

9. APPENDIX

1		2		3		4		5		6		7	
Feature	*VIF*	*Feature*	*VIF*	*Feature*	*VIF*	*Feature*	*VIF*	*Feature*	*VIF*	*Feature*	*VIF*	*feature*	*VIF*
Hours	9.511482	Hours	44.754547	Hours	22.537304	Hours	12.199393	Hours	7.460438	**Hours**	**6.683903**	DayofWeek	4.443352
DayofWeek	1.344301	DayofWeek	6.609376	DayofWeek	6.528261	DayofWeek	6.471841	DayofWeek	6.431702	DayofWeek	6.430474	FlagEvent	2.205457
FlagEvent	2.200521	FlagEvent	2.304799	FlagEvent	2.294235	FlagEvent	2.270668	FlagEvent	2.261955	FlagEvent	2.261913	FlagMorning	2.286293
FlagMorning	77.705863	FlagMorning	14.250235	FlagMorning	6.331573	FlagMorning	4.426439	FlagMorning	2.985724	FlagMorning	2.797675	FlagAfternoon	3.778921
FlagAfternoon	95.607768	FlagAfternoon	7.352102	FlagAfternoon	5.065315	FlagAfternoon	4.277836	FlagAfternoon	3.902833	FlagAfternoon	3.850508	FlagWeekend	2.007280
FlagEvening	**96.094758**	FlagWeekend	2.067593	FlagWeekend	2.063474	FlagWeekend	2.020570	FlagWeekend	2.016906	FlagWeekend	2.016904	FlagRushHour	1.434310
FlagWeekend	1.476856	FlagRushHour	1.548116	FlagRushHour	1.517761	FlagRushHour	1.517403	FlagRushHour	1.516681	FlagRushHour	1.516327	FlagEventPeak	1.480061
FlagRushHour	1.222883	FlagEventPeak	1.482754	FlagEventPeak	1.482670	FlagEventPeak	1.482284	FlagEventPeak	1.481165	FlagEventPeak	1.481164	CountGiraStops	1.216771
FlagEventPeak	1.453866	CountGiraStops	1.218897	CountGiraStops	1.218835	CountGiraStops	1.217509	CountGiraStops	1.217453	CountGiraStops	1.217047	MarathonPathFlag	1.245485
CountGiraStops	1.129297	MarathonPathFlag	1.254166	MarathonPathFlag	1.252826	MarathonPathFlag	1.252517	MarathonPathFlag	1.252303	MarathonPathFlag	1.247015	CountJams	1.294459
MarathonPathFlag	1.135839	CountJams	1.308006	CountJams	1.307990	CountJams	1.307987	CountJams	1.307649	CountJams	1.298245	MeanSpeed	1.388774
CountJams	1.225431	MeanSpeed	1.450020	MeanSpeed	1.446461	MeanSpeed	1.445669	MeanSpeed	1.443879	MeanSpeed	1.443327	CountTrainStops	1.155119
MeanSpeed	1.231878	CountTrainStops	1.164586	CountTrainStops	1.164473	CountTrainStops	1.163837	CountTrainStops	1.163682	CountTrainStops	1.156681	CountMetroStops	1.250789
CountTrainStops	1.128762	CountMetroStops	1.293773	CountMetroStops	1.293332	CountMetroStops	1.293229	CountMetroStops	1.293204	CountMetroStops	1.251768	Precipitation	1.233781
CountMetroStops	1.230773	Temperature	55.956426	**Temperature**	**49.122892**	Precipitation	1.284625	Precipitation	1.233847	Precipitation	1.233781	Sun	3.267458
Temperature	1.871199	**Humidity**	**70.728666**	Precipitation	1.285476	Sun	3.297862	Sun	3.291119	Sun	3.290570	WindSpeed	4.262293
Humidity	2.328853	Precipitation	1.298561	Sun	3.410316	WindSpeed	4.912856	WindSpeed	4.757197	WindSpeed	4.757159	CarreirasCount	2.844132
Precipitation	1.198152	Sun	5.116139	WindSpeed	6.003969	WindCardinal	12.540124	CarreirasCount	5.538024	CarreirasCount	3.194288		
Sun	3.181818	WindSpeed	6.062103	WindCardinal	14.154756	**CarreirasCount**	**5.538552**	**StopsCount**	**8.625465**				
WindSpeed	1.662512	WindCardinal	14.615578	CarreirasCount	5.541643	StopsCount	8.635630						
WindCardinal	1.400387	CarreirasCount	5.541718	StopsCount	8.644580								
CarreirasCount	2.262074	StopsCount	8.670817										
StopsCount	2.232699												

Table 8 – VIF test on base dataset

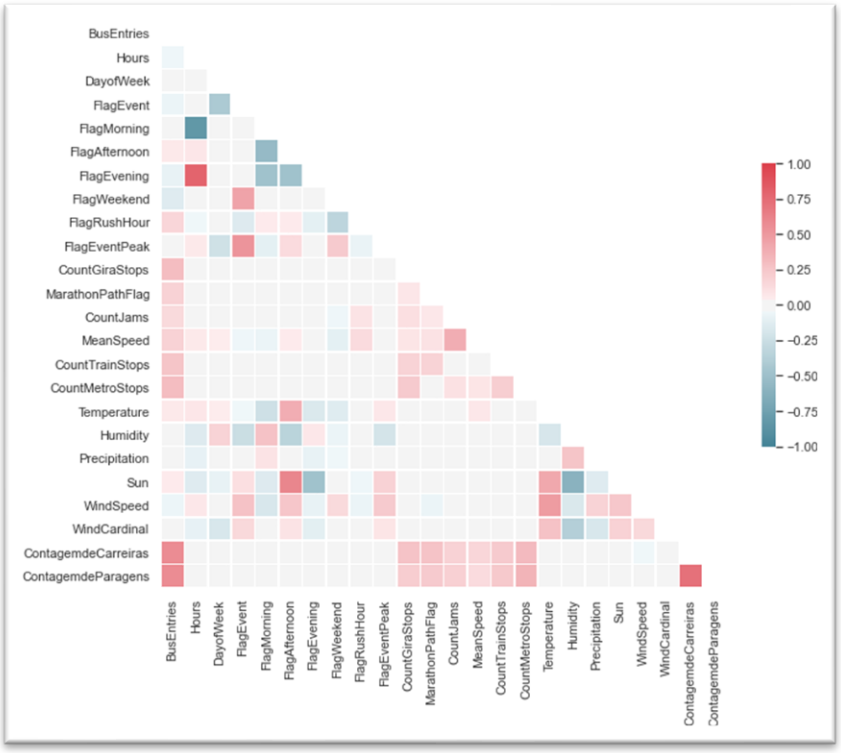

Figure 23 – Base data set Correlation Matrix