Universidade do Minho
Escola de Engenharia
Departamento de Informática

Tiago Rafael da Fonseca Fontes

Bike-sharing docking stations identification using clustering methods in Lisbon city

December 2021

Universidade do Minho
Escola de Engenharia
Departamento de Informática

Tiago Rafael da Fonseca Fontes

Bike-sharing docking stations identification using clustering methods in Lisbon city

Master pre dissertation report
Integrated Master's in Informatics Engineering

Dissertation supervised by
Paulo Novais

December 2021

COPYRIGHT AND TERMS OF USE FOR THIRD PARTY WORK

ACKNOWLEDGEMENTS

After a long journey of working, thinking, researching and pushing myself to new challenges, the end has come. For not being alone in any single moment, I have to thank my family and my girlfriend, my friends, professor Paulo Novais and all professors in my academic path, CEiiA and, in an ever grateful way, Miguel Arantes.

Thank you, family, Dad Joaquim and Mom Rosário for all these incredible months of belief, faith and fantastic support. For me, you are magnificent persons that give me affection and love a single day to keep me happy and strong. It was not easy in several moments: sometimes we regret our choices, we doubt the process and the path we have taken. At the end of the day, you were always there for me, especially for listening to. More than half of this achievement is yours because you have always kept the stability and care to focus when needed. I will be forever grateful.

There are only a few words, but... I have to especially thank you, Inês, my sweet, dear, patient life partner, for believing in me. For being the light in the foggy nights. Whenever I thought about turning around, you have made everything clear: the whole work and effort I have made to get here, the close I was every time I doubt myself and the emotional support you have given me... there are no words for thanking all you have done for me. My success is, of course, yours.

Thank you, my friends, for being trustful and excellent listeners in this long path we have crossed, for the support, and all the words I take from you. We have shared the same difficulties, feelings, and smiles and cheers when we have taken a step forward. You have made such a difference in my life for sharing those moments and experiences that I am deeply thankful for sharing one of the most important cycles in my life with you. Your optimist and positive energy had a vital contribution to this journey. Most of you started as schoolmates, but I take some of you as my friends for life.

Special thanks to you, Professor Paulo Novais, for being so patient when dealing with different moods and emotional states, helping me keep focused and motivated. I have to sincerely thank you for being there all this time, even when I did not reach for help. Just knowing you were there eased my mind and made things less complicated. You made me push myself for new challenges and goals I had never thought I would reach, keeping me concentrated and trying to go further as possible in my tasks. For everything, a huge thank you.

I would also like to thank CEiiA, the company that has believed me since the beginning. Since the first day I contacted, the support has been incredible. Special Thank you to the Data Science team, the ones I had the pleasure to work with. Everyone made me feel amazing, helping and cheering with me all the barriers we have faced and turned over. All this work is also yours, and I thank you for all the advice, suggestions and learning you have passed me. Affectionately, I would like to thank you, Paulo Figueiredo, for all your contributions and teaching me. You are an essential pillar in this journey, and as I have always said to you, you have a special place in my heart.

Exceptionally and gratefully, I thank you, Miguel Arantes. You have been an incredible supervisor, friend, confidant, and most importantly, a great human being who came every time I needed you. You have guided me through the most challenging times and, for that reason, I will be forever grateful for everything you have done for me.

I sincerely thank you all and will be forever in your debt.

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity.

I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

ABSTRACT

Urban mobility has enormous impacts on environmental, economic and, social levels, promoting necessary eco-friendly means of sustainable transportation. Soft mobility (specially bike-sharing services) plays a crucial role in these initiatives since it provides an alternative for hydrocarbon fuel vehicles inside the cities. However, choosing the best location to install soft mobility docks can be difficult since many variables should be considered (e.g. proximity to bike paths, points of interest, transportation access hubs, schools, etc.).

In Portugal and Lisbon's most specific case, we have been facing critical traffic jams and congestion, being the city elected in recent years as the "most congested city in the Iberian Peninsula". For that reason, and being the most populated city in the country, the natural pendular movements in the capital turns it a spotlight regarding mobility and sustainability.

After realising the real-world problem and the necessity of solving the pollution and negative impacts of carbon emissions solutions, this thesis aims to understand the mobility patterns within site, looking to optimise the current bicycle sharing system implemented in the study area and promote carbon-free solutions with the most efficiency as possible.

To take this task into practice, one of the most accurate ways of characterising the location of citizens is mobile data from personal cellphones, which can provide critical information regarding demographic rate, traffic trajectories, origin/destination points, etc., and can aid in the installation of soft mobility platforms, as explained in the document.

This dissertation presents a decision support system to study existing and new bike-sharing docking stations, using mobile data and clustering techniques for three Lisbon council parishes: *Beato*, *Marvila* and *Parque das Nações*. Throughout the followed methodology, it was possible to compare the findings with the in-site docking stations from the public bike-sharing system and discover and suggest new points for installing these soft-mobility solutions. In outr perspective, through the implementation of this study it will be possible to increase using rate concerning carbon-free transport modes in the city.

In summary, the work plan followed and the system implemented is viable to be replicated in other contexts and cities since mobility issues are well known in several locations worldwide. By taking advantage of similar device mapping, it would be possible to apply the system to other locations and cities and potentially reduce carbon emissions.

KEYWORDS Smart Soft mobility, Real Use Case, Lisbon, Soft Mobility, Artificial Intelligence, Sustainability, Forecasting

RESUMO

A mobilidade urbana tem enormes impactos nos níveis ambiental, económico e social, promovendo os meios de transporte sustentáveis necessários e amigos do ambiente. A mobilidade suave (especialmente os serviços de partilha de bicicletas) desempenham um papel crucial nestas iniciativas, uma vez que proporciona uma alternativa sustentável aos veículos emissores de carbon dentro das cidades. No entanto, a escolha do melhor local para a instalação de docas de mobilidade suave pode ser difícil, uma vez que muitas variáveis devem ser consideradas (por exemplo, a proximidade de ciclovias, pontos de interesse, centros de acesso ao transporte, escolas, etc.).

Em Portugal, e tendo em conta o caso específico de Lisboa, temos enfrentado engarrafamentos e congestionamentos críticos, tendo sido a cidade eleita nos últimos anos como "a cidade mais congestionada da Península Ibérica". Por essa razão, e sendo a cidade mais populosa do país, os movimentos pendulares naturais na capital tornam-na um foco de atenção no que diz respeito à mobilidade e sustentabilidade.

Depois de compreender o problema, de escala global e real, bem como a necessidade de resolver a poluição e os impactos negativos das soluções smissoras de carbono, esta tese visa compreender os padrões de mobilidade na ciade mencionada, procurando optimizar o actual sistema de partilha de bicicletas implementado na área de estudo e promover soluções livres de carbono com a maior eficiência possível.

Em termos práticos, uma das formas mais precisas de caracterizar a localização dos cidadãos são os dados móveis de telemóveis pessoais, que podem fornecer informações sensíveis relativamente à taxa demográfica, trajectórias de tráfego, pontos de origem/destino, etc., que podem ajudar na instalação de plataformas de mobilidade suave, como explicado ao longo do documento.

Assim, esta dissertação apresenta um sistema de apoio à decisão para estudar tanto as estações de bicicletas partilhadas existentes, bem como identificar as novas estações, utilizando para tal dados móveis e técnicas de *clustering* (agrupamento), em três freguesias do concelho de Lisboa: *Beato*, *Marvila* e *Parque das Nações*. Através da metodologia seguida, foi possível comparar os resultados com as estações de bicicletas relativas ao sistema de partilha de bicicletas públicas, bem como encontrar e sugerir novos pontos para a instalação destas soluções de mobilidade suave. Na nossa perspectiva, este estudo, se implementado, conduzirá a um potencial aumento da taxa de utilização de modos de transporte sem carbono na cidade.

Em suma, o sistema implementado é viável para ser replicado noutros contextos e cidades, uma vez que as preocupações em torno da mobilidade são bem conhecidas em vários locais em todo o mundo. Tirando partido de um mapeamento de dispositivos similar, será possível aplicar sistema noutras localizações e potencialmente reduzir emissões de carbono.

PALAVRAS-CHAVE Mobilidade Suave Inteligente, Caso Uso Real, Lisboa, Mobilidade Suave, Inteligência Articial, Sustentabilidade, Previsão

CONTENTS

I INTRODUCTION

II STATE OF THE ART

III MODEL DEVELOPMENT

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AI	Artificial Intelligence
AoI	Area of Interest
CML	*Câmara Municipal de Lisboa*
CEPT	European Conference of Postal and Telecommunications
CP	*Comboios de Portugal*
BSS	Bicycle Sharing System
DCAI	Distributed Computing and Artificial Intelligence
DNN	Deep Neural Network
EC	European Commission
EU	European Union
EEP	European Environment Pact
ETSI	European Telecommunications Standards Institute
EMEL	*Empresa Municipal de Mobilidade e Estacionamento de Lisboa*
ESA	European Space Agency
GDPR	General Data Protection Regulation
GIS	Geographic Information System
GPS	Global Positioning System
GSM	Global System for Mobile Communication
ICT	Information and Communication Technologies
KNN	K-Nearest Neighbors
LMA	Lisbon Metropolitan Area
MDPI	Multidisciplinary Digital Publishing Institute
ML	Machine Learning
NMT	Non Motorised Transport
PAM	Partitioning Around Medoids
PoI	Point of Interest
SDG	Sustainable Development Goals
SSE	Sum of Squared Errors
INE	*Instituto Nacional de Estatística*
UN	United Nations
UAV	Unmanned Aerial Vehicle
WWTP	Waste-Water Treatment Plant

Part I

INTRODUCTION

1

INTRODUCTION

1.1 SCOPE OF WORK

Currently, increasing population growth has created several concerns, especially in urban centres, as it brings new environmental, economic, and social challenges, as shown by Singh et al. (2017). Climate change and environmental degradation represent not only a European but also a global concern.

The United Nations (2015) Agenda for Sustainable Development, adopted by all Member States of the United Nations (2015), provides a standard blueprint for peace and prosperity for people and the planet, now and for the future. The seventeen Sustainable Development Goals (SDGs) are at its heart, an urgent call to action from all countries - developed and developing - in a global partnership. They recognise, among several topics, stimulating economic growth - all while tackling climate change. Thirteenth on the list of seventeen is "Making cities and human settlements inclusive, safe, resilient and sustainable", clarifying the global need to promote sustainability in one of the most fundamental ways, mobility.

In this sense, aiming to urge communities and cities to adopt sustainable habits regarding climate changes effects, the European Commission (2020) (EC) has released the European Environment Pact (EEP), which is the plan outlined to define a strategy that enables a modern economy, efficient in the use of resources, and competitive. This plan aims to boost the efficient use of resources through the transition to a cleaner economy and reduce pollution. Mobility plays a crucial role in sustainability, as it is one of the main focuses of the EEP. Lam and Head (2012) summarise sustainable urban mobility as "the ease, convenience, accessibility and low cost in terms of travel in these centres, seeking minimum impact on the environment and other inhabitants". Hence, the latter contemplates obvious objectives in this sense, such as investing in environmentally friendly technologies and implementing cleaner and healthier public and private transport forms.

Portugal, in particular, as we documented by Climate ADAPT (2021), has driven several initiatives aimed at promoting a less polluted environment concerning mobility, encouraging the purchase of electric vehicles, limiting the circulation of cars in certain areas of some cities, increasing bicycle lanes, increasing shared vehicles in urban centres, among others.

Portugal also has some examples of its implementation in terms of sustainable technologies, which are documented and studied by Liberato (2018), as is the case of Aveiro and Lisbon, as stated by Portuguese American Journal (2021), in terms of the soft mobility modes implemented, taking advantage of Bicycle Sharing

Systems (SBP). However, implementing these soft mobility solutions carries high costs, as seen by DeMaio (2009).

As demonstrated by Macioszek et al. (2020), Jan Wisniewski - RESETI (2019)) or Winslow and Mont (2019), the examples of cities such as Warsaw, Paris, Barcelona, among others, there is a need to plan the location of these solutions to maximise their efficiency and, ideally, make them the mode of transport chosen by citizens when it comes to their efficiency to mobility. Considering the high concentration of people and flow of resources in urban centres, it is essential to face these displacements as a serious sustainability challenge.

Typically, cities are more developed in the digitalisation component, promoting Information and Communication Technologies (ICT) Mary K. Pratt - TechTarget (2019) mechanisms in the enhancement of efficiency and quality of operations and services. Thus, in recent years a concept has been created to define these urban centres. A *smart city* Albino et al. (2015) is an urban area that leverages digital innovations, tools and organisational principles in order to help the community evolve to become more sustainable, inclusive, prosperous and creative, ultimately benefiting the individual citizen. Lisbon is thus already considered a *smart city* by Smar50Awards (2019), so by taking advantage of these intelligent ecosystems, it is possible to collect accurate information that will allow extracting and generating knowledge about them, serving as a basis for the management of goods, resources and services efficiently.

1.2 MOTIVATION

Apart from the massive side effects from air pollution and climate changes which affects the planet every year, the European Environment Agency (2021) alerts the communities, especially the European ones, for air pollution being "the biggest environmental health risk in Europe". As stated by the mentioned source, the "levels of air pollutants still exceed EU standards and the most stringent World Health Organization guidelines"

Moreover, *Lisbon* in particular, has been a special point of attention in Europe concerning air pollution and private car usage, as highlighted by News (2017), being elected as "the most congested city in the Iberian Peninsula" in recent year.

Without compromising our capacity to move, we intend to minimise the environmental impacts associated with it. To this end, it is pertinent to strengthen the participation of public transport and **soft modes** in the modal split, the adoption of technologies that increase the energy and environmental efficiency of transports or the development of soft mobility, i.e., forms of travel under their own steam (cycling, walking, scooters, roller skaters, for example.), always aiming at more sustainable mobility.

Mobility contemplates several forms and several ways of being realised. However, as seen in Section 1.1, one way to increase sustainability is by using softer modes, particularly bicycles. However, this particular model also brings challenges in rebalancing bicycle systems (taking bicycles from overloaded stations to free stations), weather conditions, city relief, among other factors, are challenges in implementing these soft modes of transport.

Nevertheless, the need of reducing the use of private transport modes such as car has already been world-wide spread, due to the necessity of a global and immediate change in order to reduce the already negatives effects in the planet. According to INE - Instituto Nacional Estatística (2019), in partnership with the EC, showed that in 2016, the private car share rate was 83% of the passenger per kilometre modal split, while bus and train did not exceed 9% and 8%, respectively. More specifically, in the Member States, the percentage of car use in land passenger transport varied between 69% in Hungary, 89% in Portugal and 90% in Lithuania. Although this study was only concerned with road and rail choices, we can easily understand from these results that the preference for the car as a means of locomotion is a risk factor both in terms of pollution and sustainability, but also in terms of mobility as a whole, causing congestion and travel difficulties. Being sure that all modes of transport have their role in the transport system, there is a need to pursue policies that ensure more intelligent mobility and, consequently, more sustainable mobility.

In *Lisbon*, as presented in Figure 1, public transport modes share shows an important change that should be adopted all over the world, since it is still far from the perfect values.

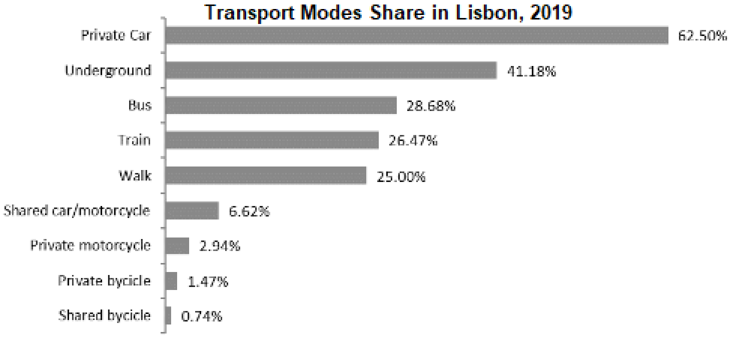


Figure 1: Transport Modes share in *Lisbon*, 2019.

Source: Bernardo et al. (2019)

Figure 1 was taken from Bernardo et al. (2019) study regarding *Lisbon*, and it represents the most common transport modes used in the usual trip home-school/work. As it shows, private car usage in *Lisbon* is considerably lower than the mean value in Portugal, 89%. This is a crucial indicator of the investment and concern regarding mobility in the Portuguese capital. Hence, it is vital to take into consideration other essential indicators concerning other transport modes. Train, bus and underground represent the most used options apart from the private car, highlighting the significance of these public transport modes. However, as was also concluded in the referred study, "smart mobility solutions are yet not well disseminated compared to the traditional solutions".

1.3 MAIN GOALS

Nowadays, given the importance of mobility and the numerous displacements that can be registered in urban centres, we realise that there is a clear need to understand and study the populations' behaviours in terms of the displacements. As mentioned in Section 1.2, it is necessary to develop mechanisms so that the populations can have access valid, safe and appropriate options regarding mobility, so that when it comes to making decisions, the most sustainable choices possible are made.

In partnership with CEiiA, a Centre of Engineering and Product Development designs, implements, and operates innovative products and systems. CEiiA operates in several exciting domains such as aeronautics, mobility, ocean, space and automotive. Corroborating the importance of CEiiA in the mobility context is the presence of this entity in COP26 (explained here by nationalgrid (2021) and Climate Change Conference (2021)),as shown in CEiiA (2021).

Hence, it is proposed in this project, the design and development of a **decision support system** in the context of urban mobility, that based on accurate data extracted from mobile phones, can **help** in the **territorial planning** of the **mobility points** of the city of **Lisbon**, in order to **maximise** the **use** of these modes, promoting **greater** use of the **existing means** or **introducing** them in subareas that currently lack these solutions.

Using techniques from the sphere of Artificial Intelligence, such as *Machine Learning* and *Clustering* techniques, as well as the use of analytical and other AI-related models, it will be possible to model mobility realistically, thus allowing the determination of the most central points to increase the existing mobility services, contributing to the diversification of sustainable solutions and of course, increase the quality of life of the inhabitants.

In this way, the objectives of this dissertation are based on:

- Understanding the different existing mobility solutions, their specifications and limitations
- Analyse the study area and existing mobility infrastructures to restrict the focus of action and associated context
- Study and interpretation of mobile data provided in order to identify mobility patterns and implications on soft mobility solutions
- Comprehend existing bicycle sharing system implemented in the study area
- Development of clustering models in the scope of urban mobility in order to identify new bike-sharing docking stations
- Analyse and explore the obtained results, proceeding to their interpretation and examining their real impact in terms of mobility planning

As we have seen so far, mobility represents a massive focus for all stakeholders. With technological advances and the passing of the years, where the use of own vehicles has prevailed, it is necessary to adapt and improve mobility implementation policies in major urban centres. In cities where there are no options for soft mobility, these mechanisms should be implemented given already their advantages presented, and in those where they are available, these mechanisms should be improved to provide suited conditions for inhabitants and tourists.

1.3.1 *The problem*

Thus, this dissertation intends to take an **actual case study**, the city of *Lisbon*, where it will seek to develop work plan based on **clustering** and **prediction** mechanisms to understand and mitigate the main problems related to soft mobility based on **accurate data**.

Thus, the main obstacles addressed in the literature, and as seen in more detail in Chapter 2, are:

- Establishing the Spatio-temporal relationships between Bicycle Sharing Systems (BSS) stations, points of interest (POIs), population clusters, etc.
- Assess the impact of these services or the lack of them in terms of mobility
- Quantity of data and direct influence on more traditional models
- To model and study the **variations** of **demand** for **soft** modes of transport, taking into account external factors

1.3.2 *Challenges*

Regarding the challenges imposed by the problem at hand, the use of a **actual case study**, the data that will be made available by a **telecommunications operator**, providing "accurate" information regarding mobile phones status, also known as Global System for Mobile Communications (GSM) data. Hence, an additional precaution and preliminary analysis of them is necessary since they may contain some inaccuracies, impermissible values, etc., as mentioned in Section 2.5, in this context, due to communication failures, data processing blockages, etc. As described further in Section 2.6, especially in the area of project development in the scope of ML, it is pretty essential to perform a preparatory phase and preprocess the data so that they are clean and with quality to catapult good results in the implementation phase of the models.

Thus, as described by Check Point (2021), GSM data, formerly known as "Groupe Special Mobile," is a **world-wide standard** for **digital wireless mobile phones**. The standard was originated by the European Conference of Postal and Telecommunications Administrations (CEPT) and further developed by the European Telecommunications Standards Institute (ETSI) as a standard for European mobile phones, with the intention of developing an open, non-proprietary standard for adoption world-wide. It has been remarkably successful, with more than one billion people using GSM phones as of early 2004. Hence, the telecommunication operator provided some information regarding cellphone activities from the GSM datalake, concerning the month of January 2020, containing information regarding mobile phone activity, including the temporal reference for each sample, and of course, the geolocation associated.

In this sense, and as much of the data will be relative to geographical positions, the use of the Global Positioning System (GPS) will be a significant factor. For that reason, it is important to sum up some of the most concerns regarding the GPS signal and in which cases can it be affected from external factors, since this study **relies**, in particular, **geographical information** of each activity registered. Additionally, other constraints are covered further, in Section 5.2.

Hence, according to GPS.gov (2021), satellites transmit their signals in space with a sure accuracy, but what you receive depends on additional factors, including the geometry of the satellite, signal blockages, atmospheric conditions, and the characteristics/quality of the *design* of the receiver. The same source refers that, "GPS-enabled smartphones are typically accurate to within **4.9 metres** with clear skies. However, their accuracy worsens near buildings, bridges, and trees." Some devices can have higher GPS accuracy with dual-frequency receivers, depending of course, on the built-in hardware.

There are several causes for the **error** associated with the GPS signal, some of them explored and described by Taczanowska et al. (2008). According to the authors, some of which are:

- Atmospheric interference.
- Calculation and rounding errors.
- Ephemeris (orbital path).
- Multi-path effects -> signal reflection on buildings, walls or other obstacles.

Stated the importance of knowing the physical limitations of the geographical representation used in this study, it is important to illustrate some of the main side effects that might occur when using GPS signal for geographical representation. Thus, one of the main causes of GPS inaccurate signal broadcast is real world interference, such as signals reflected off buildings, etc. Figure 2 illustrates how this kind of limitation affects signal's quality.

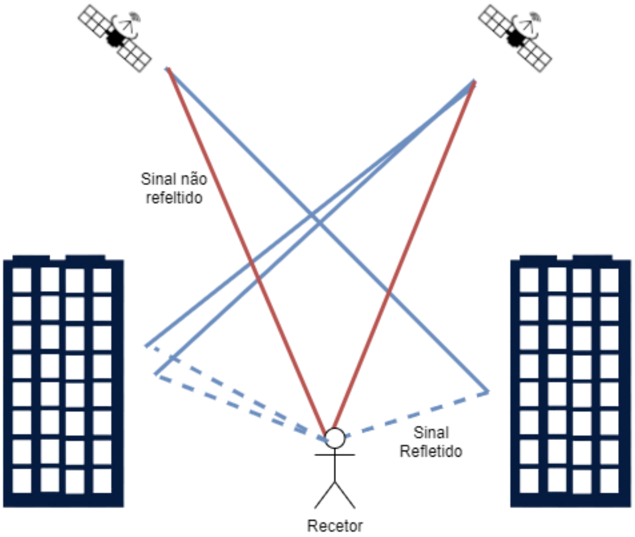


Figure 2: Example of a GPS error: signal reflections on buildings

Adapted from: Taczanowska et al. (2008)

Less systematically, radio interference, large solar storms and even satellite maintenance periods can create mismatches between the data. In this sense, it will be **essential** to **consider** the **accuracy** and **error associated** with the **GPS signal** as a **challenging factor** in this **dissertation** precisely because it is **not entirely accurate** and may consequently **influence possible conclusions** drawn from the data analysis.

Another factor that will be a limitation will be some of the **data** used in the realisation of the dissertation, namely those that will come from a telecommunications operator. In current times, data disclosure is particularly legislated and limited due to the **General Data Protection Regulation** - available at European Union (2016) - and made by the European Parliament and the Commission of the European Union. This is a European law regulation on privacy and personal data protection, applicable to all individuals in the European Union and European Economic Area created in 2018. It also regulates the export of personal data outside the European Union and European Economic Area. The GDPR aims to give citizens and residents ways to control their data and unify the European regulatory framework. The regulation was adopted on 15 April 2016. After a two-year transition period, has been officially in operation since 25 May 2018.

As it is a Regulation, this is directly applicable to the 28 Member States, without any transposition for each jurisdiction. And for the context of this thesis, it is important to highlight the **implication** of **such guidelines**. The regulation introduced significant changes to data protection rules, imposing new obligations to organisations whose breach is punishable by heavy fines that could amount to up to 4% of the global annual turnover of up to 20 million euros.

The mentioned Diploma clarifies the concept of **personal data** and result from it rights for data subjects, such as the **right to data portability**, the **right to be forgotten**,etc. So, considering these kind of limitations and new concepts concerning data regulation, the rules for obtaining the **consent** of the data holders become much more demanding. Alongside these constraints, another novelty with GDPR is the introduction of new principles and concepts that should guide the processing of data such as **pseudonymisation**.

In this sense, the data provided by the telecommunication operator is **aggregated**, since it will not be admissible to monitor and analyse the journeys made by users individually. In short, considering the **error associated** with the **GPS signal** and the **spatial aggregations performed**, these will undoubtedly be considerations to be taken into account, so they should be faced as challenges in the study of **mobility** since the study of traffic flows will have to be based on these **geographical limitations**.

1.4 WORK SCHEDULE

A work schedule for the dissertation's admission has been created to provide an estimate of the time required to meet the various objectives that make up the dissertation during the academic year 2020/2021. This timeline will contribute to planning and efficient time management for the conclusion of this dissertation. After researching the dissertation topic and analysis of the objectives to be achieved, eleven tasks were identified. For completing the various tasks, twelve months was estimated, from October 2020 to October 2021. In order to achieve the previously mentioned objectives, the following schedule for the dissertation is presented.

- Understanding the challenges inherent to mobility
- Bibliographic research
- Identify relevant methodologies

- Preparation and design of an implementation model
- Data validation and exploratory analysis
- Typification of perspectives and analytical approaches in mobility
- Characterisation of mobility approaches
- Methodology development
- Validation of the obtained results
- Model implementation
- Code review and final-tuning
- Dissertation writing
- Preparation for public dissertation's presentation

As a way to better illustrate the planning performed, a Gantt chart is presented below.

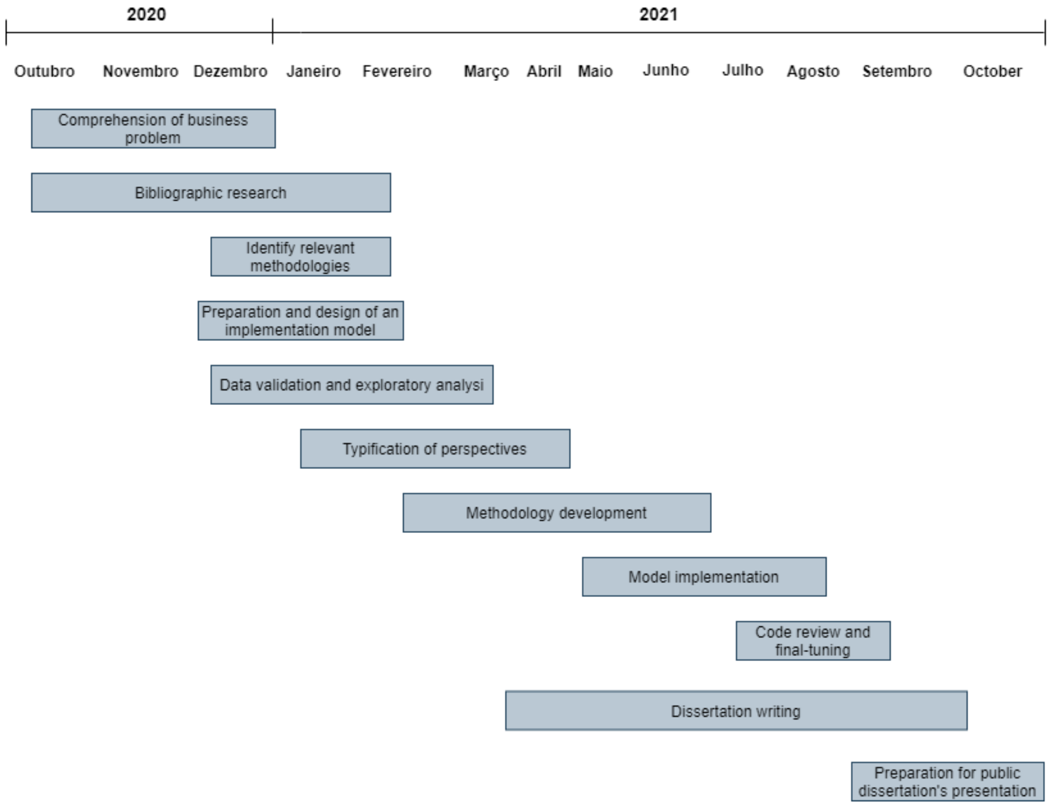


Figure 3: Work schedule initially proposed for the current dissertation.

1.5 DOCUMENT STRUCTURE

The dissertation report has the following structure:

Chapter 1 is an introductory chapter dedicated to the contextualisation, motivation and objectives of the topic, thus introducing some basic concepts for the presentation of the dissertation and where the goals are defined and briefly dissected.

The 2nd Chapter aims to present state of art concerning mobility and Artificial Intelligence (AI). This Chapter will address the AI concept, its importance and the different levels of clustering that exist. Then, it will present some notions about Artificial Intelligence concepts, Machine Learning and also the Clustering notion and interesting algorithms for this case study. Additionally, mobility and its different solutions regarding soft modes and other important information is provided.

Chapter 3 aims to describe the main characteristics regarding the current mobility solutions in Lisbon city, with a particular interest in the study of our area of study and the available Points of Interest(PoI) that might affect mobility patterns, as studied throughout the document.

Chapter 4 pretends to provide a deeper analysis of the GSM data gathered from the telecommunication operator in order to scrutinise visual patterns on mobility choices, compare different time frames across the three parishes in our study area, etc. This will allow a more accurate interpretation and discussion of results in the following chapters since it will provide an extensive analysis of data and identification of agglomerations, interesting areas and so on.

Chapter 5 describes, in detail, the major data transformations, algorithms applied and a temporal sequence of steps that have been taken. This is a very important part of the current document, once it distinguishes this study from state of the art found, but also narrates the key decisions and operations made. Also, the Chapter there are extensively expressed the main constraints and limitations faced.

The 6th Chapter will seek to detail the results obtained, as well as to look for real-world interpretations regarding mobility context and some interference of external factors such as demographically, socially or economically. There is a deeper analysis through some evaluation metrics in the AI, and in particular in Machine Learning (ML), but also mobility analysis considering the key indicators defined.

Chapter 7 summarises the work developed in the previous chapters, leaving the future work ahead but also making stock on the goals set, discussing the main achievements and the potential improvements that could be considered. Additionally, there are summarised the scientific contributions that have an outcome from the realisation of this thesis, and also a brief reflection and sharing experience moment about the CEiiA's professional experience, the team inclusion insights, among others.

Part II

STATE OF THE ART

2

STATE OF THE ART

The way we move from one place to another efficiently and responsibly is one of humanity's biggest challenges nowadays, as reported by European Commission (2020). In this millennium, several initiatives seek to improve the efficiency of movements, especially in large cities. As interest in the study of mobility has increased, several tools have been developed in this area, often using modelation as the primary technique for its study. This is due to models' characteristics allows the study of mobility and other import contexts, being essential in scientific and technological advances regarding the ability to move.

The importance of Artificial Intelligence in the mobility context has been discussed by Michael Majster et al. (2021), mentioning "There is still huge potential for AI to help solve many critical transportation and mobility challenges now and in the future, both in terms of improving effectiveness and efficiency" . For that reason, AI have been linked to mobility in recent years, being some of those contributions for such a big and global issue described by Maayan, Gilad (2020), IAV (2021) and Akker, van den (2021), for example.

2.1 ARTIFICIAL INTELLIGENCE

In the early 1950s, Alan Turing, a famous mathematician, raised the question "Can Machines Think?" in his research paper called "Computing Machinery and Intelligence", becoming the basis of the goal of Artificial Intelligence, as described by Krach et al. (2008). However, some philosophers in the scientific community debate this issue appeared in the early 17th century, as documented in Turing (2009).

We can then say that **Artificial Intelligence** is the area of computer science that deals with the ability to make machines intelligent. As is common knowledge, computers typically execute instructions provided by humans, which are called programs. However, this new way of promoting intelligence machines in an "autonomous way" to solve problems, we call Artificial Intelligence, as Haenlein and Kaplan (2019) dubbed.

2.1.1 *Machine Learning*

Arthur Lee Samuel is one of the pioneers in the field of applied machine learning, as in McCarthy and Feigenbaum (1990). It was in 1952 that he built a program to play checkers. This program was able to memorise the moves that had occurred and combine this information with values from an objective function. For Samuel, *Machine Learning* is defined as "the field of science that gives computers the ability to learn without being explicitly programmed".

Very briefly, we can define *Machine Learning* as the subset of Artificial Intelligence that gives the ability to learn and improve automatically through past experiences.

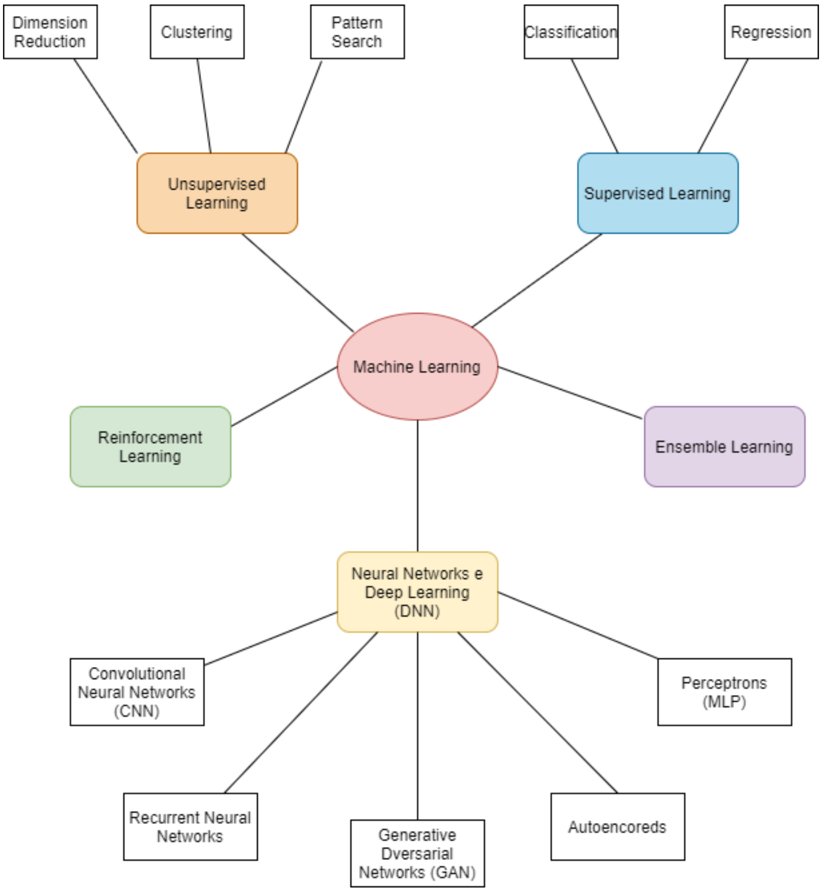


Figure 4: Main areas of Machine Learning

Adapted from: Patel (2018)

Figure 4 illustrates the several areas that are linked to the ML concept,and also some of their context of application, such as classification, regression, pattern search, among others.

Thus, *Machine Learning* models can be classified into three main and distinct categories, as mentioned by Telikani et al. (2021) or Sah (2020), among several available sources:

1. **Unsupervised Learning -** this category, called unsupervised learning, in contrast to the previous one, in the training process of these models, no value of *output* is provided to the system. Thus, the models will have to organise and group the data in an "intelligent" and meaningful way. In this strand, the problems of **clustering**, **dimension reduction** and **pattern search** are presented.

2. **Supervised Learning** - *input* data is sent to the models, with the intention that the machine learns the expected result for this same data, i.e., perceives which *output* should be returned. In these situations we speak of *labeled data*, that is, when the input data is accompanied by the final result (*output*). Within this strand, we find applications such as **classification** and **regression**.

   This approach is known for dividing the dataset into three distinct parts to feed the following three phases:

   - **Training** - the model is fed with data in order to find patterns in it that are necessary to recognise the given outcome (*label*). Throughout this process, the parameters are updated since there is a *loss* function that will penalise poor predictions. This is a repetitive process, which leads to improvements in performance.
   - **Validation** - in this phase occurs an evaluation of the generalisation ability of the model since it will be confronted with new data. Data will be sent without the associated *label* so that the model itself can identify the result. The evaluation will be done by the *loss* function mentioned .
   - **Test** - The goal of this phase is to verify the model performance, after the training process, with data without *label*. This is a crucial stage because the model's actual generalisation capacity and consequent aptitude for the problem with which it is being confronted are certified.

3. **Reinforcement Learning** - this type of learning is based on a trial-error principle. Briefly, the algorithm will rely on a function that will return positive rewards when the direction of the action taken is positive, and when the action brings negative effects, a penalty will be assigned.

2.1.2 *Clustering*

As stated previously, Clustering is one of the most important areas of study regarding Artificial Intelligence and Machine Learning applications. As stated by Bano and Khan (2018), data clustering if one of the most essential, common and interesting task to classification of patterns in different areas, including data mining, pattern recognition, artificial intelligence, among others.

This procedure, as the name suggests, is based on making **cluster** of **entities** ton based on **their similar features.** For that reason, when using clustering techniques, we aim to create high quality clusters, looking for high intra-class similarity and low inter-class similarity, being that one of the most premise in this context.

In order to create similar clusters a distance measured is used. However, there are several distance measures applied: Euclidean distance, Manhattan or taxicab distance, Mahalanobis distance, etc.

There are many data clustering algorithms, being them usually categorised into two groups: **unsupervised linear** and **unsupervised non-linear.**

Considering the multiple application areas of machine learning and different approaches in this scope, it is possible to classify Clustering procedures into the following categories:

- Hierarchical Procedures
- Partitioning Procedures

- Density-based Procedures
- Model-based Clustering Procedures
- Grid-based Procedures
- Fuzzy Clustering

For that reason, being these algorithms widely applied and very assimilar in some aspects, a deeply description of some of these categories is going to be presented, in order to better comprehend differences and important advantages of each category and main algorithms.

Clustering Based on Hierarchy

This category is a paradigm of cluster analysis to generate a sequence of nested partitions (clusters) which can be visualised as a tree or so to say a hierarchy of clusters known as cluster dendrogram. Hierarchical trees can provide a view of data at different levels of abstraction. The basic idea is to construct hierarchical relationship among data in order to cluster their points.

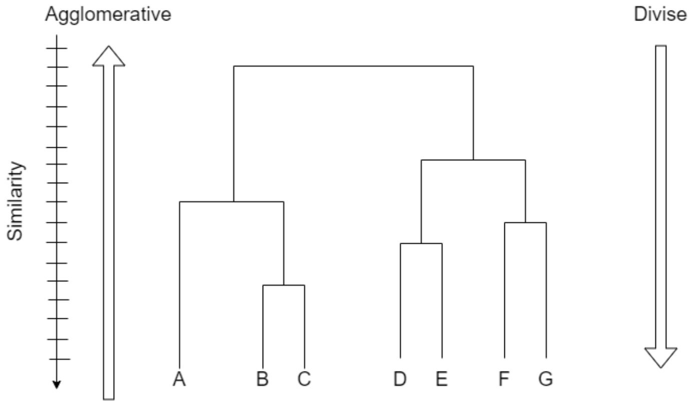


Figure 5: Hierarchical clustering behaviour

Adapted from: Giordani et al. (2020)

Figure 5 illustrates the process that hierarchical clustering algorithms follow, in a high level. Hence, as Figure suggests, this method seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:

- **Agglomerative**: This is a "bottom-up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- **Divisive**: This is a "top-down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

Some typical algorithms of this kind of clustering include BIRCH, CURE, ROCK, Chameleon. As described by Xu and Tian (2015), BIRCH realises the clustering result by constructing the feature tree of clustering, CF tree, which will dynamically grow when a new data point comes. Regarding CURE, suitable for large-scale clustering, takes random sampling technique to cluster sample, offering robustness against noise and outliers. This algorithm introduces a novel data structure which is capable of summarising the information that is maintained about a cluster. Thus, it is possible to work with compress data and develop multiple nodes. These nodes work as tiny clusters and depict the summary of original data.

Regarding ROCK, this is another agglomerative hierarchical algorithm, which works based on the concept of "links". "Links" are used for measuring proximity between a pair of data points with categorical attributes Kaufman and Rousseeuw (2009). A criteria function is then calculated considering the measure of how proximate or similar clusters are, in order to evaluate their "goodness".

Considering CURE, it was developed for identifying more complex cluster shapes, following Frey and Dueck (2007). Instead of using a single centroid it assumes many separate fixed points as clusters and a fragment m is used to shrink these diverse points towards centroids. Moreover, these scattered points after shrinking represent the cluster at each iteration and the pair of clusters with the closest representatives are merged together.

Chameleon, in instance, is capable of "measuring the similarity of two clusters based on a dynamic model" Thambusamy and T (2011). One important factor to consider is its merging criteria, more detailed in comparison to CURE. Essentially, the process occurs in two phases: firstly create a graph, which contains links between each point and its N-nearest neighbour. Secondly, the mentioned graph is recursively split by a partitioning algorithm, resulting in many tiny unconnected graphs. The merging process occurs during the second phase, when each sub graph is considered as an initial sub cluster, merging them if the resultant cluster has inter connectivity and closeness to the two parent clusters prior to merging.

Clustering Based on Partitioning

Described the principles regarding hierarchical approach, it is important to focus now other clustering procedure, **Partitional clustering**. This is a very disimilar approach when compared to hierarchical one, since the first yields an incremental level of clusters with iterative fusions or divisions, and **partitional**, as the name suggests, assigns a **set** of **data points** into **K clusters** with no hierarchical structure, as in Everitt et al. (2011).

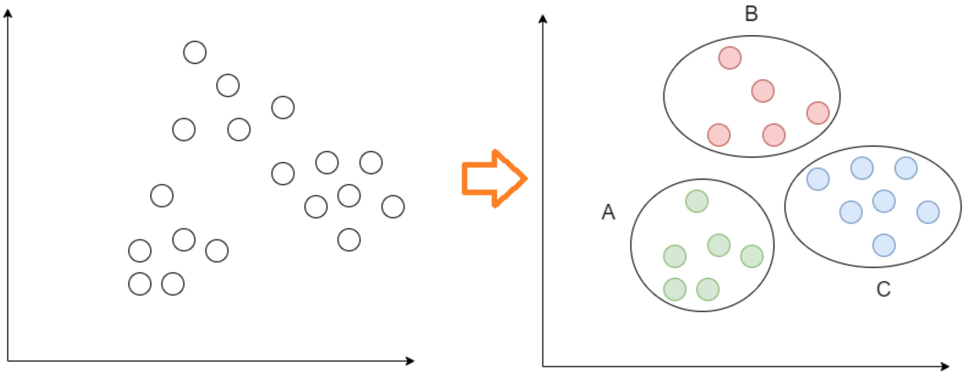


Figure 6: Partitioning clustering behaviour

Adapted from: Jin and Han (2011)

As Figure 6, the initial set of points were assign to the K value, i.e., the number of clusters that are supposed to be output. In This case, as an illustrative, it was used K = 3, creating three distinct groups.

One of the most popular algorithms from partitioning clustering is **K-Means**,as studied by Likas et al. (2003). It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimise the sum of distances between the data point and their corresponding clusters.

In a brief way, the k-means clustering algorithm mainly performs two tasks:

1. Determines the best value for K center points or centroids by an iterative process.
2. Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

In order to complete these two tasks mentioned above, we can describe the K-Means as following, as mentioned by Sharma (2019):

1. Step 1: Choose the number of clusters k
2. Select k random points from the data as centroids
3. Assign all the points to the closest cluster centroid
4. Recompute the centroids of newly formed clusters
5. Repeat steps 3 and 4

However, this steps must take into account a **stopping criteria**, since these steps are must come to and end and provide an output. In this sense, and following the mentioned source, the next stopping criteria are applied to define the final result:

- Centroids of newly formed clusters do not change
- Points remain in the same cluster
- Maximum number of iterations are reached

As mentioned, K-Means really takes advantage of the notion of **centroid**. As described by Kumar (2020), "the centroid point is the point that represents its cluster". Hence, Centroid point is the average of all the points in the set and will change in each step and will be computed by:

$$C_i = \frac{1}{||S_i||} \sum_{x_j \in S_i} x_j$$

For the above equation:

- C_i: i'th Centroid
- S_i: All points belonging to set_i with centroid as C_i
- x_{-j} : j'th point from the set
- $||s_{-}i||$: number of points in set_i

The idea of the K-Means algorithm is to find k centroid points (C_1, C_2, . . . C_k) by **minimising** the **sum** over each **cluster** of the sum of the **square** of the **distance between** the **point** and its **centroid**. Hence, it is pretended to get the **lowest** values for the metric of **Sum** of **Squared Error** (SSE), which is calculated as presented below:

$$C_1, C_2, \cdots, C_k = \operatorname{argmin} \sum_{i=1}^{k} \sum_{x \in S_i} ||x - C_i||^2$$

After analysing the behaviour of the K-Means algorithm, there is also a very similar algorithm that comes to mind: the **K-Medoids.** This is a very related algorithm in terms of approach and steps to be applied during the clusters identification process, but there is one of **main distinct feature** from K-Medoids and K-Means (for example): the **centroid process identification**. In K-Means, the **final centroids** are not the **actual point** but the **mean of points present in that cluster.**

However, in K-Medoids this cluster identification process is a bit different since the purpose of K-Medoids clustering is to make the **final centroids** as **actual data-points**, i.e., the identified centroids **belong** to the initial dataset, and most specific, to the cluster group they represent.

Moreover, the algorithm of K-Medoids clustering is called Partitioning Around Medoids (PAM), as mentioned by Jin and Han (2010). For that reason, using the PAM, the process of identification of cluster becomes slight different, resulting in differences in the cluster identification, as explained in Figure 7.

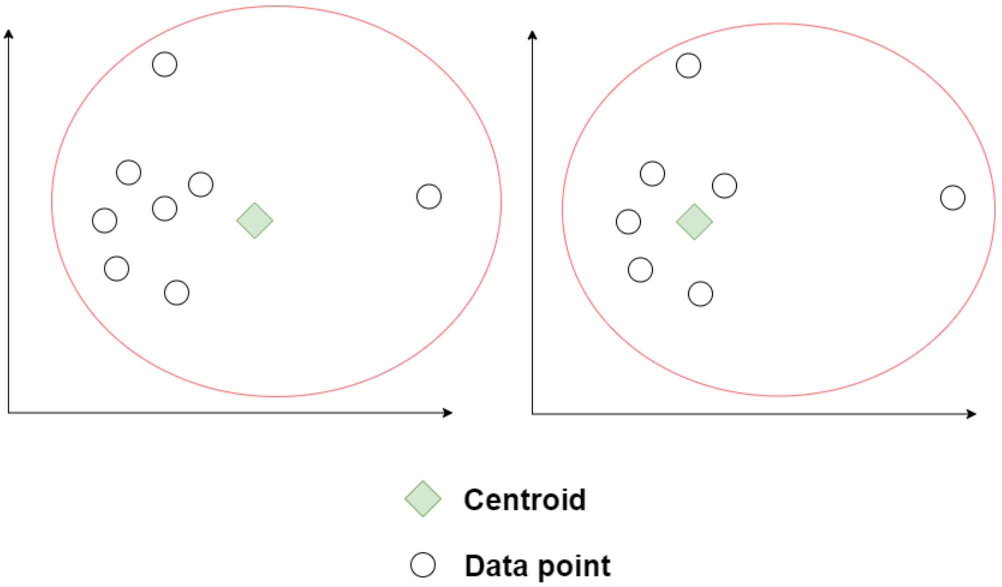


Figure 7: K-Means vd K-Medoids centroid output

Adapted from: Jin and Han (2010)

From Figure 7 is possible to notice the **different influence** that an **outlier** produces using the K-Means and the K-Medoids. This is an **important** factor to take into account, since it is also noticed in other partitioning algorithms, such as **CLARANS**.

CLARANS means Clustering Large Applications based on RANdomized Search, which is another popular partitioning clustering algorithm. K-Medoids clustering technique can resolve the limitation of the K-Means algorithm of being adversely affected by noise/outliers in the input data. But K-Medoids proves to be a computationally costly method for considerably large values of 'k' (number of clusters) and large datasets, which is also another constraint.

The CLARA algorithm was introduced as an extension of K-Medoids, as mentioned by Shiledarbaxi (2021). It uses only random samples of the input data (instead of the entire dataset) and computes the best medoids in those samples. It thus works better than K-Medoids for crowded datasets. However, the algorithm may give wrong clustering results if one or more sampled medoids are away from the actual best medoids.

CLARANS algorithm takes care of the cons of both K-Medoids and CLARA algorithms besides dealing with difficult-to-handle data mining data, i.e. spatial data. It maintains a balance between the computational cost and the influence of data sampling on clusters' formation.

As mentioned by Dagli (2019), the CLARANS algorithm follows the process presented next:

1. First, it randomly selects k objects in the data set as the current medoids. It then randomly selects a current medoid x and an object y that is not one of the current medoids.
2. Then it checks for the following condition: *Can replacing x by y improve the absolute-error criterion?*
3. If yes, the replacement is made. CLARANS conducts such a randomised search l times. The set of the current medoids after the l steps is considered a local optimum.
4. CLARANS repeats this randomised process m times and returns the best local optimal as the final result.

Then, the CLARANS repeats this randomised process m times and returns the best local optimal as the final result.

Comparing the three mentioned algorithms, and as mentioned by Gandhi and Srivastava (2014) , Table 1 sums up the main aspects for analysis.

Table 1: Comparison of K-Means, K-Medoids and CLARANS

Parameters	K-Means	K-Medoids	CLARANS
Complexity	O (i k n)	O (i k (n - k)^2)	O (n^2)
Efficiency	Comparatively more	Comparatively less	Comparatively more
Implementation	Easy	Complicated	Complicated
Sensitive to Outliers?	Yes	No	No
Advance specification of No. of clusters 'K'	Required	Required	Required
Does initial partition affects result and runtime?	Yes	Yes	Yes
Optimised for	Separated clusters	Separated clusters, small dataset	Separated clusters, large dataset

Source: Gandhi and Srivastava (2014)

As we can see in Table 1, the three mentioned algorithms differ in several points, such as complexity, outliers influence and its implementation difficulty. Nevertheless, they keep a common feature, needing the **initial value** of K to be provided. Hence, it is presented next a popular approach to identify this initial value.

The **Elbow method**, as mentioned by Yuan and Yang (2019), is a well-known approach to determine the "optimal" number of clusters into which the data may be clustered. The method consists of plotting the explained variation as a function of the number of clusters, and **picking elbow** of the **curve** (the point of inflexion on the curve) as the number of clusters to use. That means that the elbow of the curve represents the **sweet spot** where the **error** does not decrease in a significant way that still represents a benefit in the increase of K value.

Figure 8 illustrates the elbow method application.

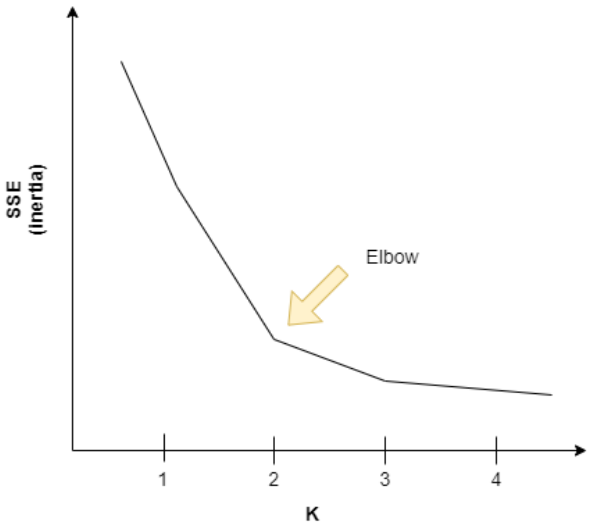


Figure 8: Illustration of Elbow method: K = 2 is the "optimal" value

Adapted from: Yuan and Yang (2019)

As we can see in Figure 8, the elbow method provides a graphical analysis to figure out the "optimal" value for K. In this illustrative scenario, the elbow occurs when K = 2. Nevertheless, this heuristic does not provide any guarantee of selecting the best K value, since these algorithms depend on the problem/project that are being applied, and can suffer some modifications. One of them is increase the value for K, a bit higher than the elbow inflexion, if the context of the problem benefits from it.

Clustering Based on Density

Lastly, it now presented the **Density-based** clustering techniques. As mentioned by Sander (2010), Density-Based Clustering refers to unsupervised learning methods that identify distinctive groups/clusters in the data, based on the idea that a cluster in a data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density. The data points in the separating regions of low point density are typically considered noise/outliers.

Thus, compared to centroid-based clustering like k-means, density-based clustering works by identifying “dense” clusters of points, allowing it to learn clusters of arbitrary shape and identify outliers in the data.

One of the most popular algorithms in density-based clustering context is **DBSCAN**. This algorithm, as mentioned by Svirca (2020), estimates the density by counting the number of points in a fixed-radius neighbourhood or ϵ and deem that two points are connected only if they lie within each other’s neighbourhood. So this algorithm uses two parameters, being them:

1. ϵ - denotes the Eps-neighborhood of a point

2. MinPts - denotes the minimum points in an Eps neighbourhood

For that reason, DBSCAN needs these **two** parameters, as initial configuration. For that reason, the algorithm, using a high level description, can be reported by the following steps:

1. Arbitrary select a point P.
2. Retrieve all point density reachable from P with regard to ϵ and MinPts.
3. If P is a core point a cluster is formed.
4. If P is a border point, then there is no point that is density-reachable and DBSCAN moves to the next point.
5. This process is continued until all the points are processed.

Thus, giving as input N objects to be clusters and global parameters ϵ and MinPts, the desired and expected output is clusters of objects. An illustrative example of this algorithm is shown in Figure 9.

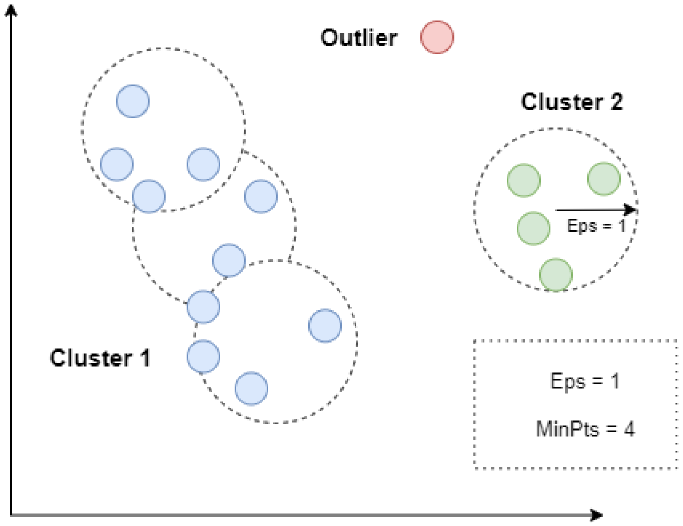


Figure 9: Illustrative example of Density-based clustering algorithms process

Adapted from: Wang et al. (2015)

As described in Figure 9, the algorithm takes advantage from the two initial parameters, being them ϵ and MinPts, and is applied to the initial dataset and build the clusters following these configuration values.

According to Wang et al. (2015), DBSCAN presents some interesting **advantages**, such as find arbitrary shaped clusters using MinPts parameters, the order of the point in the database is insensitive and also, it can handle noise and outliers due to the way the clusters are calculated. However, the main **disadvantages** presented are facing more difficult to perform well with large differences in densities and also being not suitable when various density involve.

2.2 CLUSTERING AND OTHER APPROACHES IN MOBILITY CONTEXT

As mentioned at the beginning of the current Chapter, AI has becoming more and more useful to solve real-world problems, and in specific, regarding mobility context. In fact, the use of **clustering methods** in mobility issues is becoming also very helpful in territorial planning and transportation management.

One of this examples is presented by Lee et al. (2017), who developed work in order to capture dynamically changing neighbourhoods based on two different types of urban mobility data. Through clustering temporal urban mobility signatures of alternative transportation users in Washington, D.C., this work provides implications about the characteristics of different types of mobility data and research directions.

To complete this research, two alternative transportation datasets were used, with the purpose of better understanding the role that human mobility plays in differentiating regions or neighbourhoods within a city. Specifically, we compare car share dataset and also a **bike share** datasets by examining the spatiotemporal usage patterns (signatures) within Washington. Through cluster analysis, we group regions of the city that behave similarly in their temporal mobility patterns.

In specific, the authors described their work, that consisted on counting the numbers of cars parked and bikes rented from within each hexagon region. After that, they have applied a transformation process to each polygon, reaching a normalised vector of more than 400 features, which represented a combination of weekday and weekend signatures. Then, a clustering method have been applied to the hexagons, using **K-means clustering** algorithm.

According to the authors, their initial findings suggest that it is possible to extract temporal patterns within these data, and furthermore can help to delineate regions within a city.

Regarding other implementations in order to aid in urban mobility planning and bicycle-sharing systems management, other approaches have been applied, as stated by Martinez et al. (2012). This study, as noted by the authors, aims to design and deploy of a bike-sharing system developed for *Lisbon*, using a non AI approach to solve it. The mentioned study **does not focus** in the **same area** of *Lisbon* as this dissertation does (since the mentioned study focus on a central area of *Lisbon* city), this provides a useful and interesting approach to solve the bicycle-sharing system management.

The system have taken into consideration several variables such as capacity of the docking stations and the fleet costs. They followed a mathematical approach and, apart form the results obtained, the authors highlight that "e key issues identified in the literature as the main drawback of this type of systems were confirmed, showing that solving the vandalism and the theft problem of the system infrastructure are the main challenge of bike-sharing programs."

Overall, they call attention that planning and manage these kind of systems must take into account the **demand** and **costs**, but also other external factors such as the **vandalism**.

2.3 MOBILITY AND SOFT MOBILITY

Soft mobility includes all forms of non-motorised transport (NMT), and which therefore use only human energy to move, presented by La Rocca (2010). The development of mobility by soft modes is associated with pedestrians, bicycles and, more recently, scooters. However, considering the emerging eletrical bicycles and other "motorirised" transport modes, we are going to include here these kind of solutions, since they also fits on ""zero impact" mobility" or "active mobility", since IMT (2013-2020) highlights for the advantages and spreading of this kind of vehicles.

All the problems related to urban congestion, pollution, noise, lack of parking, and the precariousness of the urban environment, make us today see cars in a more undesirable form. We are also starting to see that soft solutions represent, in the current urban form context, significant competitive advantages over car use.

It is common knowledge that the practice of these considered modes of travel bring vital benefits, such as improved health, energy saving, reduced congestion and reduced air and noise pollution. In a less direct but still significant way, their use also results in greater efficiency of the public transport system, in numerous economic and social advantages, what we call **intermodal** uses, and in the reduction of accidents and other associated risks for the population.

Thus, the need arises to promote good conditions for pedestrians and cyclists, adopt healthier lifestyles, and more efficient urban systems, with more excellent proximity and accessibility and less harmful emissions.

Therefore, in order to guide urban public policies towards the goal of sustainable mobility protect public space, the health and well-being of citizens, it is necessary to study and understand mobility in the designated locations to promote the best conditions related to urban planning, transport, etc., so that soft modes achieve higher uses and efficiency today.

2.3.1 *Bicycle sharing systems*

The importance of primary forms of soft mobility, such as **bicycles**, are well defined and historically seen as primary solution, as reported in Zayed (2017). This simple transportation mode provided the first bicycle sharing systems (BSS), and the evolution during the last decades provides a perfect retrospective regarding these solutions. Their case studies are vast, and their efficiency is more than proven, as presented by DeMaio (2009).

Among the many success stories are cities such as Paris, Lyon, Copenhagen, Barcelona and Montreal, noted by Midgley (2011). However, even though many of these cities have already gone through the various BSS developments, the experiences of each city have allowed and still allow many others to conclude in order to implement these systems, or in case they already exist, to improve their quality and efficiency. One such example is Paris and Barcelona, who, knowing that the arrangement of stations/docks (*docks*) depends on the size and configuration of the city, have chosen to space those same docks with differences of 300m.

Two of the main factors influencing citizens' decision are **topography** and **climate**. As referenced by Sánchez-Barroso et al. (2021) and Sanmiguel-Rodríguez and Arufe-Giráldez (2019), **weather conditions** represent an essential factor in mobility in general, even more so when it comes to outdoor transport modes. In this sense, and as presented in the literature, there is a clear influence of the weather conditions on the use of BSS. Some of these systems located in Northern European regions, even close in the colder months.

On the other hand, the topography of cities represents, unequivocally, a focus of attention by users, as analysed by Matias and Virtudes (2020). Typically, users dislike **slopes** of more than 4º, stating that above 8º, they consider it impractical to use the bicycle. In that sense, lower slopes do not seem to have much influence on the choice of bicycles as a mode of locomotion, but obviously, very steep slopes will become a clear restriction to the use of this mode of transport. With this, here comes a major **problem** associated with **bike-sharing**: **unbalancing** of stations. As will be natural, users will be willing to go down the steepest inclines but will not perform the reverse route and go up. This will cause stations at higher elevations to empty, while stations at lower elevations will tend to fill more easily and possibly even overfill their capacity. With this, there have been several studies to understand the imbalance that occurs in these regions, such as that of Chiariotti et al. (2018), from hiring staff to transport bikes from the most crowded stations to the less crowded ones, or else resorting to incentives for users to use a bike at a lower level and deliver it at higher elevations, found by Singla et al. (2015).

This reality has been one of the main concerns around BSS, as they clearly affect their efficiency and often entail additional costs when choosing to allocate people and vehicles to transport bicycles to the most vacant locations. In this sense, several investigations have been made in this sense, such as Chiariotti et al. (2018) and others already mentioned, in order to optimize the process and reduce costs associated with these systems that, ideally, could be autonomously sustained by users.

As Yang et al. (2016) demonstrated, there is a need to model and understand the Spatio-temporal relationships between the different stations of the BSS, that with recourse to past data, it was possible to perceive the uses of the various stations and other indicators to apply forecasting techniques to determine usage rates according to weekdays, weekends and also holiday periods. They also noted the **importance** of the **temporal granularity**, that is, grouping in different time intervals led to different results. In addition to this evidence, data on weather conditions and the state of the weather were used.

2.4 TOOLS

The development of any *software* presupposes the use of programming tools and languages as a vehicle for implementing the system to be developed.

For different types of problems, additional tools and languages appear to be the most useful and have specific advantages and disadvantages in various aspects.

Thus, in order to **develop** an **intelligent model** of **urban mobility**, which uses clustering algorithms and Artificial Intelligence, two major languages appear immediately in this sense: **Python** and **R**, as demonstrated in the Christina Voskoglou (2017) article. In this article, Python emerges as the most popular, with 57% of users choosing this language and R with 7% in third place.

R and Python both share similar characteristics and are the most popular tools used by data scientists. Both are *open source*, with Python being a widely helpful programming language, while R is targeted at statistical analysis, as presented by Kumar, Vikash (2019).

2.4.1 *Python*

This is a developed language Python Org (2021), in the late 1980s and today is used in widely known applications such as YouTube, Instagram, Quora or even Dropbox. This is a language that owes a lot to enthusiastic developers, as mentioned by Sheromova, Vasilisa (2020) since they are the most significant contributors either to the language directly or to libraries that Python makes use of, mainly in the scope of *Machine Learning*

As said before, it is a universal language and usable in various contexts. Its main advantages are, as mentioned in Kumar, Vikash (2019):

- General use is used in different areas, and its versatility is the main reason. From *Machine Learning* to *web* development, these are some of the areas where Python has excelled.
- Learning curve - it is a language considered easy to learn, having a vast community and equally extensive documentation.
- Important libraries - besides the ease of use of the language and the simplistic syntax, Python has numerous libraries for data manipulation, collection, and control. Many of these are particularly targeted at the AI area, such as *scikit-learn*, *pandas*,etc.
- Integration - Python allows easy integration with other systems, produced on top of other languages.

However, it is also important to highlight some disadvantages of this language:

- *Threading* - due to a component called *Global Interpreter Lock* (GIL), the process of halting and Threading is complicated in this language and may even be problematic.
- Statistical Libraries - although a broadly-based language, it cannot excel in all aspects. In this case, libraries aimed at statistical studies and analysis are less developed than other areas.

And considering the scope of this dissertation project, it is necessary to consider viable alternatives for the fulfilment of all the proposed tasks with tremendous success.

2.4.2 *R*

R (2021), was developed by statisticians and basically for statisticians, as evidenced by its characteristic syntax. R contains several mathematical computations derived from statistics, so this language is highly recommended in such contexts.

Some of its main advantages, as documented for example by de Vries, Andrie , Meys, Joris, are:

- Capacity Analysis - data analysis and visualisation is an essential process in *Data Science*. R is particularly used for these development phases, enabling rapid prototyping and the ability to handle datasets within *Machine Learning*.
- Libraries and tools - R, like Python, comprises multiple libraries that help improve project performance. With the use of these packages, complexity is reduced, and greater time efficiency will be obtained, as they allow the implementation of multiple tasks in a short period.
- Suitable for exploratory work - in the early stages, where it is necessary to explore the data we have and its characteristics, or even in the process of implementing models, R facilitates these activities since it requires few lines of code to do so.

Still, and alike any tool, it has some disadvantages that should also be taken into account:

- steep learning curve - according to the scientific community, R is a challenging language and therefore entails a steep learning curve. The community, for this reason, is not as broad as Python.
- Less fast - R is considerably slower than other languages in some tasks, compared to MATLAB and Python. In some cases, libraries in R are considerably less optimised and faster than in Python.

2.4.3 *QGIS*

QGIS (2021) functions as geographic information system (GIS) software, allowing users to analyze and edit spatial information, in addition to composing and exporting graphical maps. QGIS supports both raster and vector layers; vector data is stored as either point, line, or polygon features. In common words, QGIS is one of the best tools for GIS data handling, according to AIM (2019).

Obviously, taking into consideration it is as open-source software, there a vast advantages claimed by its community. Among them, we can identify:

- Free to use - being an open-source tools, QGIS provides an important advantage when compared to other paid GIS softwares, as highlighted by GrindGIS (2019) as major positive point.
- Compatibility - since many tools are restricted to specific operating systems, QGIS is widely used due to its compatibility , being compatible with almost all Operating Systems. It works with Windows, Mac, Linux, Android, and UNIX.
- Surfeit of usable features - this tools, citing its comunity and daily users, "offers the basics and much more". For most common actions, QGIS includes useful features include vector analysis, geometry tools, sampling, and geo-processing. The easy layer fragmentation and structure offers an interesting way for visualisation and editing.
- Performance - when compared to other well-know tools in this scope, QGIS offers better processing time, versus ArcGIS, as stated by Priyanki Baruah (2019)

- Integration - QGIS allows the use of external plugins in order to improve the use of external tools, bringing new functionalities
- Community - the mirror of open-source tools is its community and developers. These people belongs and have blogs and social accounts, where new features being developed are shared and also ideas exchange, in order to contribute for the software growth.

In the other hand, some cons are identified by users, being:

- Less Beginner Friendly - To some extent, most people think that this application is confusing to beginners. This is related with menus organisation and the user experience is hard due to hidden menus and options.
- Learning curve - Due to its not obvious user interface, the learning curve in order to fully understand and know this software is long.

2.5 METHODOLOGIES

The planning and structuring of a project is a critical aspect to achieve success. In this sense, this section will address some of the successful methodologies in this kind of projects, what are their characteristics, and in the end, present the methodology that will be used to accomplish this dissertation.

2.5.1 *Popular Methodologies in Data Science*

This dissertation is inserted in the scope of the development of *software*, in partnership with the CEiiA - Centre for Product Development Engineering. In this sense, the methodology applied to this project must follow the guidelines and work methods contemplated in this institution.

CRISP-DM

Cross-Industry Standard Process Data Mining (CRISP-DM) is one of the primary methodologies regarding the development of *data science* projects. Like other development techniques, it assumes an iterative process that contemplates several stages. It is assumed that each stage may be visited as often as necessary to refine the problem understanding, data preparation and even the results stage. This evolutionary process allows for information to be exchanged between the various stages and even in subsequent iterations, as presented by Abdul Qureshi (2019).

Briefly, the stages of CRISP-DM are:

1. **Comprehension of *business problem*** - this first stage has as main purpose the definition of the objectives to be fulfilled, as well as the requirements that should be respected, taking into account a business perspective

2. **Data analysis** - the most technical component of the project begins at this stage. This is where data acquisition is carried out, followed by data exploration to understand its organisation, typology, structures, etc. Furthermore, possible problems will be identified, such as non-admissible values and balancing.

3. **Data preparation** - this step covers all the activities that are dedicated to cleaning, structuring, organising data in order to build a set of data with quality, to be later sent to the models

4. **Modelling** - in this phase, it is assumed the implementation of one or more algorithms in order to achieve the proposed objectives.

5. **Assessment** - at this stage, an analysis of the results obtained will be made, taking into account the proposed objectives and what would be the expected results. In this sense, this stage may determine the return to the *business understanding* stage in order to repeat a new iteration in case the objectives have not been fulfilled.

6. **Deployment** - this is the last phase of the development process. Thus, the end of the process typically corresponds to delivering a technical report encompassing the various phases crossed, techniques and results. In addition, a *cloud* solution may be carried out to host the various implementations.

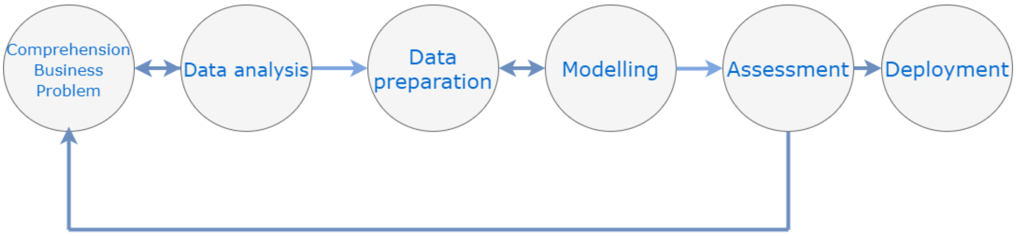


Figure 10: Lifecycle *CRISP-DM* methodology

Adapted from: Abdul Qureshi (2019)

In this way, we can state that the CRISP-DM methodology is quite popular in the development of projects like this one since it presents as its main advantage to be an **iterative process**, **very complete** and that it contemplates critical phases of *data science*. However, each one of the phases is accompanied by the writing of documentation of the state of the project.

Agile

This is a fairly common method in *software* development. This was undoubtedly a methodology that revolutionised the way we conduct projects since it allows a quick adaptation to the various circumstances during the project. In that sense, this is a philosophy that allows great iteration between people, processes and tools.

It is an iterative process and focuses essentially on refining rather than defining. These projects are characterised by flexibility and by the existence of a responsive mechanism to emerging changes, as referred to by Mike Thurber (2020).

Although there are different classifications in terms of the **cycle** of **life** of the Agile methodology, we can broadly aggregate as follows:

1. **Planning** - in this phase occur the activities of requirements gathering, problem study and conceptualisation of the system to be produced.
2. **Development** - the stage when the implementation and development take place, following the identified requirements in order to fulfil what had been previously established.
3. **Revision** - this is a critical stage because, apart from contemplating the results analysis process, it is well known for the introduction of **tests** (unitary, for example) in order to ensure the implemented functionalities.
4. **Delivery** - called *release*, this is the last stage of iteration. It corresponds to the delivery of the product in its current state.

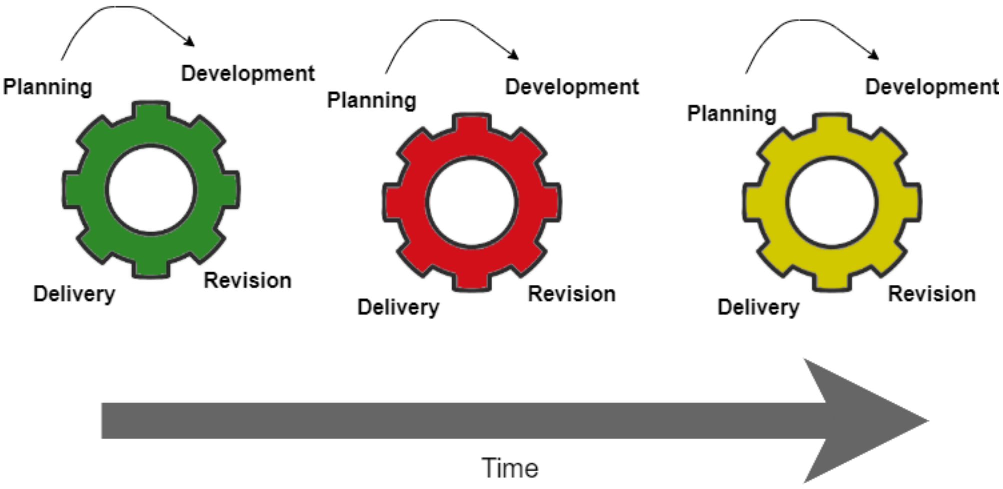


Figure 11: *Lifecycle* metodologia Agile

Adapted from: Mike Thurber (2020)

CRISPgile - CEiiA DS Team

Agile, even though it is a widely used philosophy in the *software* development industry, the characteristics of *Data Science* projects do not always facilitate the introduction of this process in their development. This is because they depend on different data, models, and adaptations between the various models and the data that will feed them.

In this sense, the use of a technique that combines CRISP-DM and Agile methodologies here called **CRISPgile** would allow us to take advantage of their advantages, allowing us to maintain an evolving and constantly documented process, combining the flexibility and adaptations that may be necessary. In short, the various stages may be characterised as follows:

- For each of the tasks, perform:
  1. **Research and Development -** in this first stage of development, a process of *research* is carried out within the scope of the topic being pursued. In this sense, the research work and study carried out can gather relevant information and move on to the development process. This second component is related to the implementation and achievement of the execution of the technical part.
  2. **Tests and Documentation -** once the development process is completed, it is necessary to code tests to the previous development process. After that, it is necessary to document all the information collected in the previous task, implementation, and results. In case the development component of the previous step is not performed, the reasons should be documented.
  3. **Review -** After the analysis, development and testing process comes the phase to review everything that was done and analyse the process again as a whole.
  4. **Delivery -** This stage includes the delivery of all the software developed, as well as the documentation that was previously written. This way, as many deliveries will be performed as there are tasks in the *backlog* (tasks list) to achieve all the objectives.

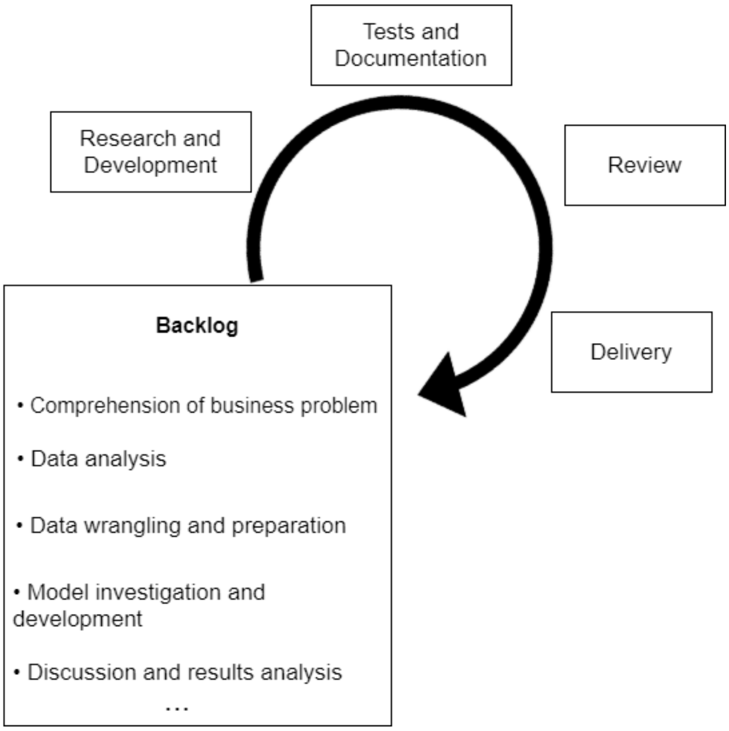

Tests and Documentation
Research and Development
Review
Delivery
Backlog
• Comprehension of business problem
• Data analysis
• Data wrangling and preparation
• Model investigation and development
• Discussion and results analysis
...

Figure 12: Proposed methodology to be applied on current thesis

The Figure 12 intends to represent the methodology that is proposed for the realisation of this dissertation. In this sense, and as it is possible to see in the Figure 12, there is a combination between the Agile method and the CRIPS-DM that was presented before. In this way, it is also an iterative process, based on several phases of CRISP-DM but with some adjustments.

Thus, it is possible to combine the Agile methodology with CRISP-DM since, for each task, an Agile cycle will be performed with the activities described above. With this, it is possible to develop the project evolutionary and iterative, so new capabilities will be added over time, and new results will be achieved.

In order to organise the various tasks towards the conclusion of the project, a **tasks** panel will be used, where the **state** of **completion** of these tasks must be indicated. For each one, a descriptive text must be applied with the following information:

```
STORY/RESEARCH

 (Given) some context

 (When) some action is carried out

 (Then) a particular set of observable
 consequences should obtain

 As a <user or stakeholder type>

 I want <some software feature>

 So that <some business value>


 ## Background

 -> Business context

 ## Technical Documentation

 -> Studies, articles, reports found

 ## How to test

 -> Commands, scripts, etc
```

With this format, it will be possible to clearly understand what the purpose of this task is and what goal it aims to fulfil. Furthermore, it will be made explicit what information has been collected in the **analysis** phase that has been carried out, the technical documentation that needs to be considered in order to understand this task and also, if the development component is included, to designate how the work carried out should be tested.

2.6 MOBILITY DATA PROCESSING MECHANISMS

During the development process in the scope of Artificial Intelligence and not only, the phase prior to the development itself is a **preparatory phase** of data, since these are the pillar of the algorithms and are the basis of the knowledge that these may acquire, as expressed by Koleva, Nancy (2019) or Haakman et al. (2020).

In this sense, the whole process of data preparation is particularly relevant, since the quality of the data is central to obtaining good results.

Thus, it will be important to address the type of data that are typically used, the decomposition in geographical terms performed in previous studies, as well as the most common problems associated with the data that will serve to feed these algorithms.

2.6.1 *Spatio-temporal data and its relations*

The datasets used in urban traffic prediction are usually characterised by spatio-temporal relationships. These datasets can be divided into three categories: static spatio-temporal, static spatial component with dynamic temporal component and also dynamic spatio-temporal, as referred to as Xie et al. (2020). Essentially, these types of datasets will differ in terms of the data that constitute them, for example whether the geographic data is related to fixed points of interest and events, or whether it is collected through mobile phone data, where there it is categorised as dynamic.

2.6.2 *Map decomposition*

The data collected on the urban context is, of course, intrinsically related to its spatial and temporal location. These data sets are often obtained through bicycles, taxis, public transport, etc. However, to proceed to their representation and measurement, it is necessary to arrange these data in relation to the time and space they characterise. One of the representations used in the literature for this purpose is the use of city map decomposition. One such decomposition method used is grid-based. For example, one such technique is presented by Zhang et al. (2016) was using a DNN - *Deep Neural Network*. i.e. Deep Neural Network, which was used as a method of traffic prediction in the city of Beijing, where the latter was partitioned into grids, considering its latitude and longitude. This is a good technique to proceed to a reliable representation of the study space.

However, another approach, though more complex, that can be used in dissection in geographical terms is to use GPS information of the vehicles, for instance, in order to map the paths taken according to the topology of the city. In this way, it will be possible to take advantage of the information of the possible and usable roads and paths, subsequently applying clustering techniques. Still, this representation technique is less convenient and simple than applying grid-based decomposition.

2.6.3 *Data concerns*

External data sources, although very convenient and easily acquired, taking into account the technological and communication advances existing mainly in urban centres, entail concerns in terms of data quality. Some of the most common problems are lack of data (*missing data*), unbalance (*data inbalance*) and also data uncertainty (*data uncertainty*). These problems are associated with the loss of performance of the algorithms, so they must be taken into consideration.

Regarding the lack of data, this is a particularly frequent scenario since, due to sensor failure or similar situations, such as bad readings, some values may not be effectively collected. Currently, the main process to mitigate this *missing data* problem is to replace these values with known ones. Lee et al. (2008) proposed a Markov factorial model to retrieve these same missing data. On the other hand, a method that also ensured excellent performance was multiple view-based method, SV-MVL from English *multi-view-based learning* that efficiently managed to fill missing data by collecting geo-sensory time series data, proposed by Yi et al. (2016).

The fact that the datasets do not present a minimally uniform distribution could greatly influence the behaviour of the algorithms. In this sense, there are several suggestions to overcome this problem, namely methods based on supervised Machine Learning algorithms such as k-Nearest Neighbors (KNN), to apply a transformation to the data, taking into account those that are found in smaller numbers, as seen by *Žižekmann* in Beckmann et al. (2015).

Another technique also present in the literature is the use of a *ensemble* method, i.e., use of combination of several algorithms, starred by Gong and Kim (2017) with recourse to *random sampling* and *ROSE sampling*.

In terms of the uncertainty relating to the data, some researchers have adopted uncertainty quantification methods to try to alleviate this fairly common problem. *Bayesian Deep Learning* is one technique that can be applied in these situations, demonstrated by Wang and Yeung (2016)

2.7 OVERVIEW ON STATE OF THE ART

During this Chapter 2, there are several topics covered, being some of them the concept of AI, clustering definition and a description of some of the most used algorithms in this AI area. After that, it was described some important notes regarding soft mobility and, in particular, the bike-sharing system's characteristics.

After that, it was explained some of the most popular tools in the Data Science problems and, in specific, ML problems. Considering the resolution of these problems, there are some standard methodologies that can be found in the literature review and were also described. Next, it was presented some of the mechanisms used in this research in order to process and handle data for mobility problems.

Going deeper by each subject mentioned, during the state of the art study it was possible to review some important clustering algorithms, especially, those who are more popular for **geopatial problems**. Among them, it was noticeable the importance of **K-Means**, **K-Medoids** and also, still less popular, the **CLARANS**. As seen previously, all these algorithms have some common and distinct characteristics. Regarding the same features among these algorithms, we can note that all of them belong to the same clustering category (partitioning clustering). This category is based on the clustering methods used to classify observations, within a data set, into multiple groups based on their similarity. However, these family of algorithms require the analyst to **specify** the **number of clusters** to be **generated**.

Thus, the initial partition affects the result and the runtime. Regarding the distinct characteristics of these algorithms we have, at least, four distinct topics: **complexity** (regarding calculations), **efficency**, **implementation ease** and **sensitivity to outliers**. Concerning the first topic, the **K-Means** algorithm is the least complex one, followed by the CLARANS and lastly, K-Medoids. This is related to the second factor, the efficiency, being K-Means and CLARANS more optimised and the K-Medoids, requiring more computational resources. Regarding the implementation, the K-Means clustering is the easier one, according to several sources mentioned previously, and for these reasons, one of the most used algorithms, especially in larger datasets. On the opposite, the K-Medoids and the CLARANS algorithms are more complicated to implement and, also because of their complexity, not so optimised for a larger dataset, even if CLARANS is still better in this field than K-Medoids.

Considering that the mentioned algorithms belong to the partitioning clustering family, and as stated before, there is a need to **provide** a the initial configuration parameter, being it, the **number of clusters** that the given dataset must be split into. For that purpose, it was found a common technique, called **Elbow Method**. The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use. Typically, these techniques used with two metrics: **distortion** or **inertia**.The first is calculated as the average of the squared distances from the cluster centres of the respective clusters. Typically, the Euclidean distance metric is used. The second one is It is the sum of squared distances of samples to their closest cluster centre, also known as the Sum of Squared Errors (SSE).

However, there are other important algorithms in the context of geospatial clustering, such as the DBSCAN. This is a popular density-based algorithm that **does not** need the initial configuration of the number of clusters, but requires two other parameters: ϵ (eps) and the minimum number of points required to form a dense region (minPts). It starts with an arbitrary starting point that has not been visited. This point's ϵ-neighborhood is retrieved, and if it contains sufficiently many points, a cluster is started. Hence, this is overall the most common techniques found in the state of the art exploration in order to solve geospatial clustering problems.

Regarding the mobility and some important concepts of ML applications, it was noticeable the use of K-means due to its ease of implementation and comparatively more efficient. In this sense, it was also presented the importance of well identify bike-sharing docking stations, especially nearby **bicycle lanes**, since users give such high importance as the docking stations are most close as possible to the bicycle lanes, considering the safety and security crucial factors for the usage of such systems.

After that, it was studied the main tools for mobility and clustering problems. In this sense, it was perceptible that Python and R are, by fast, the most used tools when it comes to ML problems and algorithm implementation. On the other hand, mobility-based problems usually suggest QGIS as one of the most important tools in this context, since it provides several features regarding vectorial operations and also visualisation. Additionally, this is a powerful tool that offers several integrations with external plugins, which increase the quality and variability of solutions.

Following this, an investigation regarding some of the common methodologies in Data Science problems was performed. Two of the most popular methodologies are Agile and Crisp-DM, the first following an iterative lifecycle and continuous delivery, while the second one represents a sequential pipeline with several steps in order to reach the delivery/deploy, at the end. At CEiiA, and as explained before, the current methodology consists in a mix of both methodologies, with particular highlight for research/investigation analysis and tests and documentation phases.

Lastly, and regarding some constraints such as GDPR and other data processing steps, the literature review revealed the common technique to use grid-maps in order to be the geographical representation for this kind of problems, since it avoids the **main restriction** of having atomic data points when working with real data, which implies the natural restrictions from the GDPR.

Part III

MODEL DEVELOPMENT

3

MOBILITY BASELINE IN LISBON

Mobility can be implemented in several forms. Nowadays, in Lisbon, we can find other alternatives to private cars regarding modes of transportation since the city is working o developing and promoting greener mobility solutions to its citizens.

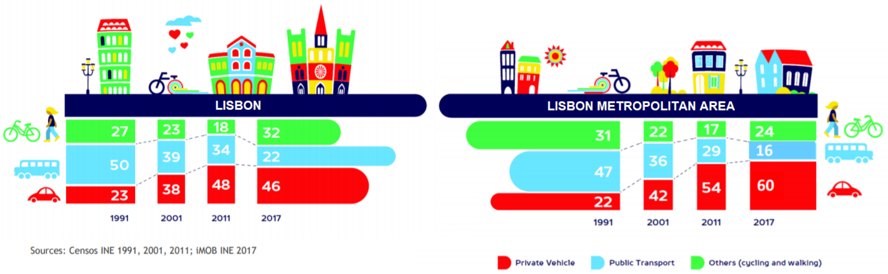


Figure 13: Lisbon and mobility development in recent years: Lisbon vs Lisbon Metropolitan Area (LMA)

Source: Machado, Pedro (2019)

According to Figure 15, Lisbon has been planning and implementing these alternative ways of moving but still noticing a worrying level of car usage.

Recognising this problem is the first step and one of the most important ones because this is going to allow us to work in a well-defined direction and with a well-determined purpose and goals.

Since we have conceptualised a new Lisbon and a new era for mobility, several advances have been reached so far and stated that a "revolution was needed according to the periodic news about climate problems and pollution". One of the priorities was public space.

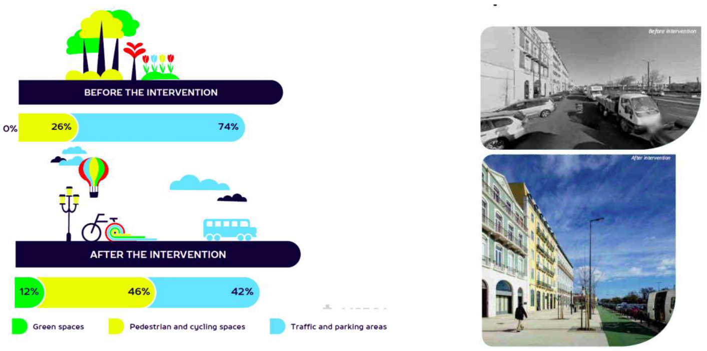


Figure 14: Evolution of the sustainable living parameters in Lisbon.

It is crucial to provide green spaces, bicycle lanes and reduce the importance and the space of vehicles in the city because we are alerting and stimulating inhabitants and all the tourists to adopt an eco-friendlier way of moving.

Since it is vital to keep progress in several subjects, *Câmara Municipal de Lisboa* (CML) (city council) has kept the investment and the promotion in mobility. After rehabilitating spaces, Lisbon created a new environment for mobility, a new era for mobility taking advantage of smartphones and smart devices. Technology and Lisbon has a very close relationship, being one of the most popular smart cities in the world, as written by Smar50Awards (2019), after being chosen to take part in Sharing Cities program, stimulated by the EU and promoting digital and technological cities across Europe.

Aiming for a sustainable future. Among several initiatives, the Lisbon council has been developing shared mobility solutions. ***Partilha Lisboa*** is one of them, looking for providing mobility shared vehicles with easier access to. Thus, the Lisbon council and some mobility providers have reached an agreement to be part of this idea. Among them, we can find **car sharing**, **moto sharing**, **bike sharing** and **e-scooter sharing**.

The LMA is since 2015 the competent authority for public inter-municipal passenger transports' services. Therefore, the organ is responsible for the mobility system's strategic objectives definition, system planning, operation organisation, supervision, financing and promotion of the services available. For that reason, LMA is responsible for managing public transport services in the municipality. In Lisbon, we can find several different passenger modes of transport services, as follows:

- Bus and Tram - *Carris* is a municipal public transport company for urban surface passenger transport: buses, trams and lifts such as: *Glória*, *Lavra*, *Bica* and *Santa Justa*;
- Metro - *Metropolitano de Lisboa* provides a public passenger transport service, in underground;
- Railway - The rail passenger transport service in urban trains in Lisbon is the responsibility of *Comboios de Portugal* (CP);
- Waterway transport - *Transtejo* and *Soflusa* provide a public river transport service integrated in the Lisbon Metropolitan Area global system;

- Taxi - The activity of rental transport in light passenger vehicles (taxis) in Lisbon is regulated by the City Council.

These are some of the well-known public transport services and providers in Lisbon. Even the focus of the current thesis is to study and develop methods to identify bike-sharing docking stations. As we saw in the literature review, these solutions are commonly integrated with other public transport services, and for that reason, it is important to consider existing transport services in this study.

3.1 GIRA - LISBON BICYCLES

Gira – Bicicletas de Lisboa is the public bike-sharing service in the city of Lisbon. The system is operated by the Lisbon Municipal Mobility (in Portuguese, EMEL) and Parking Company and was officially launched in September 2017 after a pilot phase in the parish of *Parque das Nações* under the name Lisboa Bike Sharing.

Firstly, it was only Gira operating in the urban mobility area of Lisbon, but rapidly other ways of shared mobility came up. According to EMEL (2018), at the end of 2018, 1 million trips were already taken. For this reason, being Gira a success, another micro-mobility service provides came up. Today we can find much more mobility service providers, making 8000 vehicles in total. This represents approximately 16 vehicles/1000 residents, which shows the available offers in terms of mobility. Almost 570 000 users have already been registered in at least one of the mobile platforms, making it almost 3,3 million trips done. Daily numbers aim to 20000 trips, which is a considerable number for these kinds of mobility.

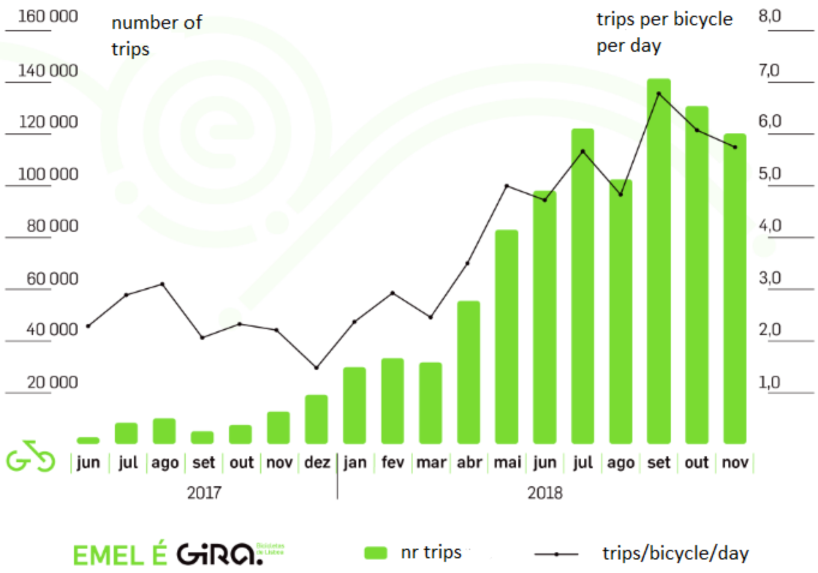


Figure 15: Evolution of GIRA bike-sharing system usage in the first years of operation

Source: EMEL (2018)

A recent study developed by Felix, Rosa and Moura, Filipe (2020) highlighted the socio-economic externalities of bike-sharing in Lisbon, analysing about 1 million trips. "This study sought to quantify the carbon emissions and polluting gases that have been avoided by people taking trips on GIRA bikes, as well as the time and money saved (compared to other modes of transport) and the medium and long-term health effects", stated by Rosa. To this end, the researchers used data from registered trips and a survey of 5000 GIRA bike users, aimed to understand whether the use of GIRA bicycles has replaced the use of another mode of transport or not.

According to the two *Técnico* researchers, the advantages of using these bikes outweigh the disadvantages. Most users prefer electric bikes to regular bikes. Almost a third of people prefer not to take the car to the city and use Gira bikes. On average, each bike ride lasts about 11 minutes.

Considering other conclusions stated by Rosa and Filipe, they claim "Globally, GIRA has reduced about 1.5 M € resulting from the reduction of atmospheric emissions, the reduction of impacts on health and road safety". "In addition to these cost savings, we should emphasize the € 42,000 in direct savings to users and the € 711,000 in time saved, compared to the other modes of transport", Rosa adds. "The study shows that 29% of people stopped using their cars and started to use a bike. 21% of people prefer to ride a bike instead of using Metro and only 12% use a bike instead of walking".

3.2 GIRA AND OTHER PUBLIC TRANSPORT SERVICES

Regarding Figure 16 , it can be seen that three mobility services are represented. Furthermore, the bus is the most used, with a fair margin when compared to the other two services. Secondly, we find the metro as the second most used out of those three. From these options, Gira is the least used service. However, we can identify two distinct areas where Gira is showing considerable relevance: *Areeiro* and *Parque das Nações*, one of the parishes from our pilot area. That said, it is important to consider *Parque das Nações* as a key place regarding mobility, but in particular soft mobility. Just about trips made across the city, it is important to figure out the most attractive places. The next graph indicates the average number of trips on the weekdays in Lisbon. The 257 trips were calculated based on the average number of trips per public transport stop per day. Also, we can analyse that the total area of the city within 1km of a public transport stop, relative to the total area of the city is over 95% and they have 2.06 stops per 1000 people.

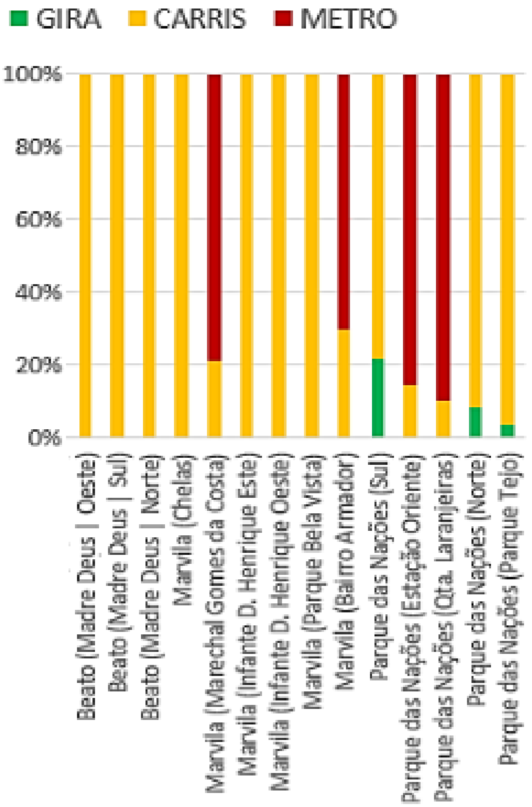


Figure 16: Cycling-bus-subway market quota (modal trip share), Lisbon.

Source: Lemonde et al. (2020)

Another relevant fact that is important to point out is the limitation of the current GIRA service in Lisbon city centre. By now, **Gira** is only **operating** in **one** of the **three council parishes** we are studying, meaning that ***Marvila*** and ***Beato*** do have not any docking station implemented for this soft-mobility operator, as we can see in Figure 17

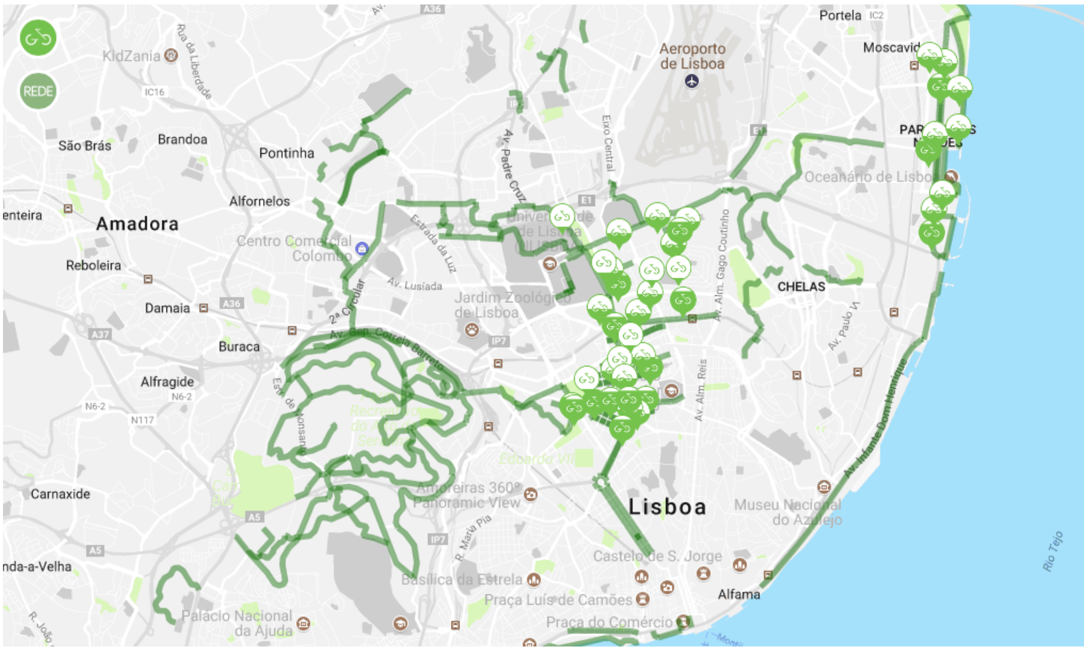


Figure 17: Current distribution of GIRA docking stations in Lisbon, 2021.

Source: GIRA (2021)

3.3 AREA OF INTEREST (AOI)

For the development of this study, several council parishes from Lisbon are considered. As highlighted in Figure 18, *Parque das Nações*, *Marvila* and *Beato* are the areas that define the spatial and geographical distribution of the work plan.

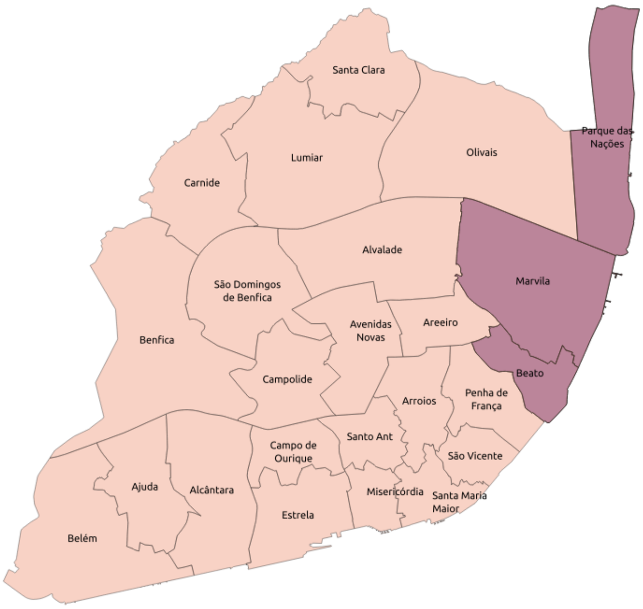


Figure 18: Area of Interest - *Parque das Nações, Marvila and Beato*.

In 2017 the investment in bicycle-sharing systems was performed for *Parque das Nações*. Nowadays, the service is expected to be available in two more council parishes (*Marvila* and *Beato*). These new locations do have several tourism access points, also having public transportation access hubs (subway, train, etc.), and new bike paths under development, as we can find in Section 3.6.

Hence, the noticeable and continuous increase in popularity of the service has allowed rapid growth of docking stations, bicycles etc. For this year, the number of docking stations and bicycles is expected to rise, following the intended investments predicted for bicycle paths for the upcoming years, as stated by Sabado (2021) and Observador (2021a).

3.4 TOPOGRAPHY ANALYSIS

Lisbon is commonly known as 'The city of seven hills, as it is characterised by irregular topography, dominated by hills and valleys of variable dimensions (Figure 19). This characteristic was the one pointed as one of the most challenging obstacles to the general adoption of soft mobility in the city, as studied by Félix et al. (2019).

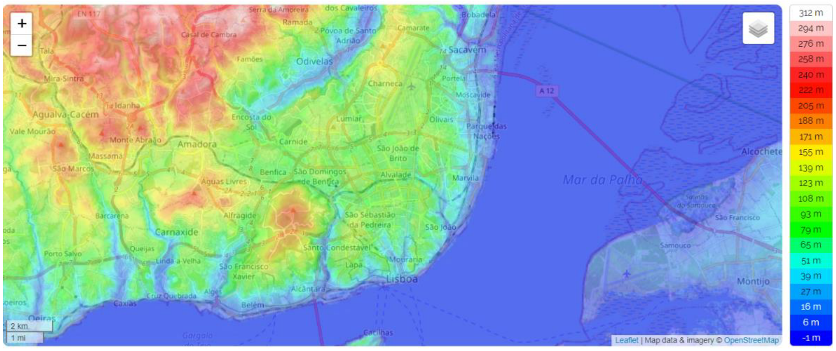


Figure 19: Representation of the Lisbon district area topography.

Source: TopographicMap (2021)

Moreover, as in most antique and medieval cities, Lisbon streets characterisation, distribution and dimension are tiny and characteristically occupied by tall buildings, minuscule sidewalks and no or reduced space to insert bike lanes or other soft-mobility dedicated infrastructure as shown in the following Figure 20. This map shows the global integration of the Lisbon street system. The warmer colours, close to red, represent parts of the city with higher values of global integration. The colder colours, even dark blue, represent parts of the city with lower values.

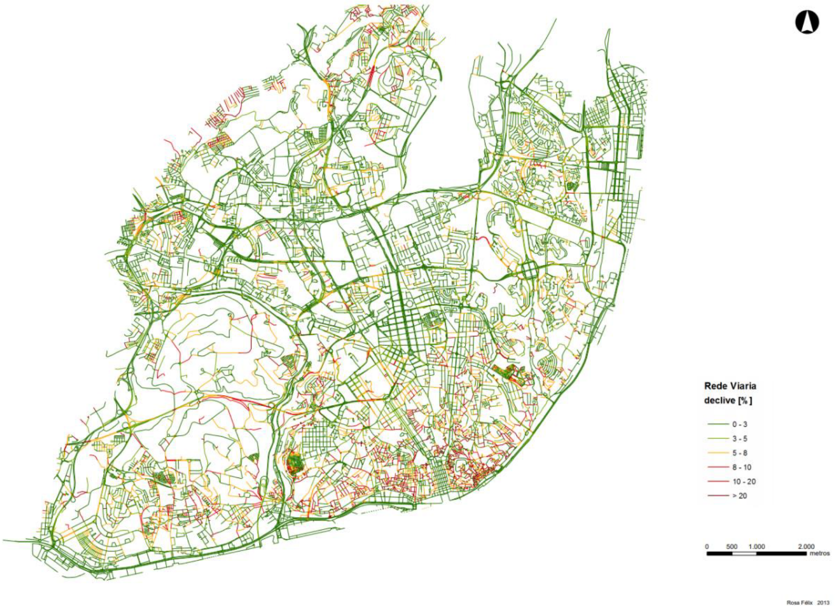


Figure 20: Lisbon axial map of the integration of the streets.

Source: UShift (2021)

Furthermore, Lisbon city centre is a historical heritage recognised by UNESCO. Thus, the central area characterised by tiny roads, buildings vertically developed represent a challenge when it comes to the innovation of the urban mobility network and to undergo any infrastructure update that affects the aesthetic and historical value of the city.

3.5 DEMOGRAPHIC ANALYSIS

One of the major factors in urban mobility is the demographic dynamics, as described by Tyrinopoulos and Antoniou (2013), such as the profile of potential users of mobility solutions. Lisbon, similarly to many other European cities, is facing a slightly decreasing population and currently is of around half a million inhabitants. However, Lisbon recently has become an attractive city for foreign and young adults due to its economic dynamism, major events such as the Web Summit, attractive weather conditions and cost of living when compared with other European capitals cities. Moreover, both national and local administrations have deployed a range of legislative initiatives to attract transnational real estate investment and new high-income residents to the country, including generous tax benefits and residency permits for large foreign investors, which will translate into an increase of foreign residents in Lisbon, especially in the city centre, as shown by Nofre and Sequera (2020).

These three parishes, located on the north bank of the Tagus river, have the specific location of being marginal areas and providing some cultural and services which are kept due to the proximity to the Tagus. Even being neighbouring council parishes, we can find some differences among them regarding population and demographic indicators — Figure 21 explicit their differences regarding the density of population indicators.

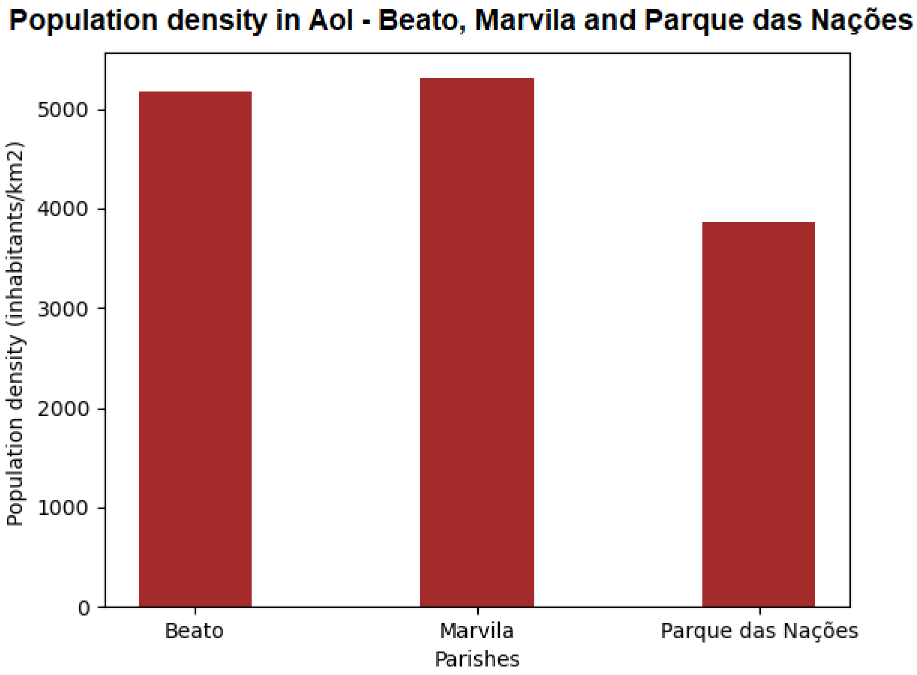


Figure 21: Population density in AoI (2011)

Data available at: INE (2011)

As shown in the plots above, the most populated region is *Marvila*. This is a parish with an area of 7,12 km2 and 37 793 inhabitants (2011), resulting in a population density of 5308 hab./km2, following the information available at CML (2011).

Parque das Nações is the second most populated area of the three explored, with an area of 5,44 km2 and 21025 residents, with a population density of 3864,9 hab/km2. The contemporary architecture of Parque das Nações, the spaces for socialising and the entire urbanisation and urban regeneration project has brought a new dynamic to the eastern part of the city of Lisbon, which in 1990 was still an industrial zone. Even if this is a mostly commercial and touristic zone, not devoted to housing, this might be an interesting parish when considering mobility demand due to the extremely relevant points of interest.

Finally, *Beato*. This is the smallest parish when compared with the two above and one of the smallest counties in Lisbon. With 2,46 km2 of extension and 12737 inhabitants, *Beato* reaches a density of 5177,6 hab/km2, considerably higher than *Parque das Nações* and close to *Marvila*. In conclusion, we could assume that *Marvila* and *Beato* are preferable regions to habit, while *Parque das Nações*, due to commercial and cultural infrastructures located there, is not a residential region that much, but potentially it is a hot spot for mobility demand. Unfortunately, there is no official data regarding population after 2011, as the last official results from National Census occurred in 2011.

However, we can define another analysis in terms of municipalities and their estimated population for the latest years. One of the most interesting points in Lisbon is being a worldwide city, allowing an intercultural concentration of people. Of course, Lisbon attracts a lot of tourists during the entire year. However, as stated before, the financial benefit for foreign and the economic dynamism associated to mild weather conditions, attracted several foreign people to settle in Lisbon.

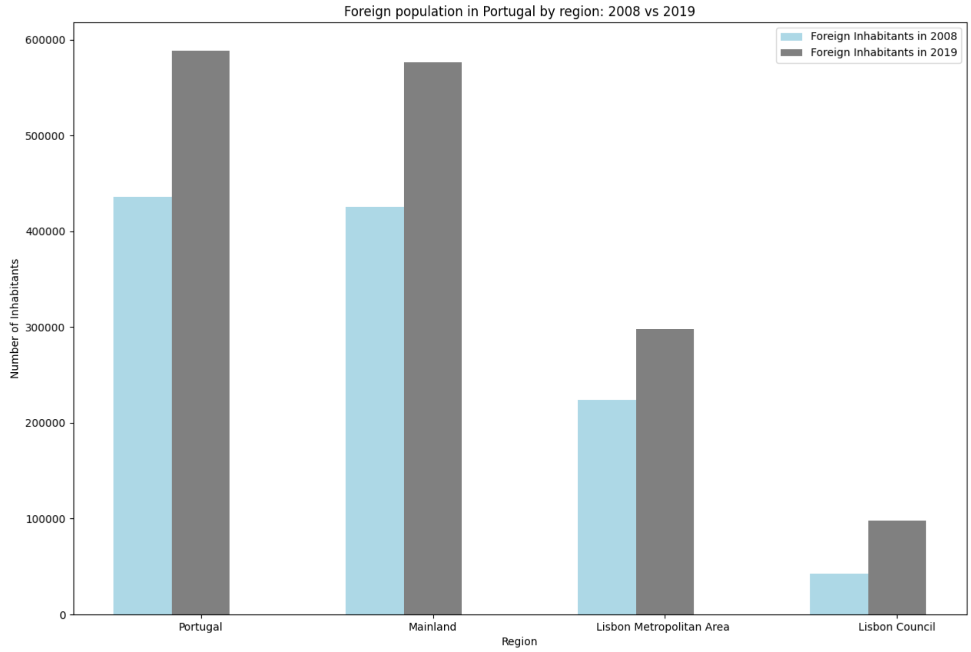


Figure 22: Foreign Population density comparison (2011).

Data available at: INE (2011)

As we can see in the bar plot above (Figure 22), from 2008 to 2019 we can see a huge difference between the several regions considered. Considering the decade between these two years, a major increase in foreign population is shown. According to INE (2011), in 2009, there were living 436020 people in Portugal, while in 2019 there were already 588976 inhabitants. As we can see in second bars plotted, compared to the first bars, this increase was about population in mainland, since the foreign population in Portugal (considering *Madeira* and *Açores* as well) is almost the same in the mainland. For that reason, we can see the LMA represents almost half of total foreign population in our country.

In 2008, there were living 224089 people in LMA, while eleven years later, there were 298175. Even though these numbers are very interesting to state the global growth of foreign people, the major increase occurred in the county of Lisbon. From 2008 to 2019, the foreign population has duplicated, with 42396 and 98151 inhabitants, respectively.

This plot notes the best way that Lisbon is attracting a lot of people who lives abroad and keep their lives in Portugal. This fact is reflected in mobility, and because choosing mobility solution has been related to cultural origins, as we are going to see further. Another interesting analysis, taking into perspective mobility, is population and its main age groups. Regarding soft mobility solutions and the middle age of users, it can be perceived as 30-40 years, as we are going to see later. This can be explained by considering the newest and the oldest groups: first, children are not the main target of these solutions. Due to responsible terms and conditions in almost all platforms, it is natural that children are not the target of shared mobility systems. Secondly, considering physical limitations that come with age, it is natural that elderly people will not take advantage of these solutions because of the effort needed.

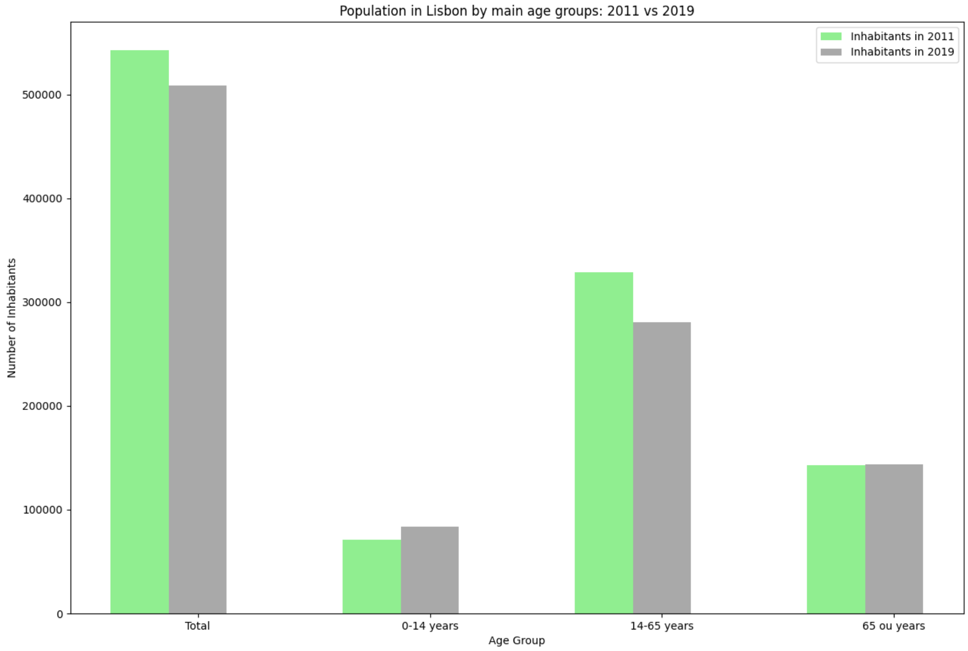


Figure 23: Different age groups population density (2011).

Data available at: INE (2011)

The population in Lisbon council has decreased from 2011 to 2019 since there were 542917 in 2011 and 2019 registered 508368 inhabitants. That said, we can figure out that the main cause for this decrease is the decrease in the group of fourteen to sixty-five years old. This can be explained by emigration, as shown in Observatorio da Emigracao (2020), with considerable high values in the last ten years, being 2013 the peak. Considering the group of zero to fourteen years old, we can identify a slight increase in this young population, where the elderly group seems to present the same distribution.

3.6 POINTS OF INTEREST ANALYSIS

In general, a Point of Interest (POI) is a specific point location that someone may find useful or interesting. Different categories can be identified, cultural buildings, services, schools or libraries, museums, among others. Therefore, these POI assume a key role in terms of mobility studying. That said, we present in this section location of some POI in the study area in order to identify potential spots of high attractiveness and understand the most candidates for moving with alternative ways of mobility. The objective is to take into consideration the POIs in the development of the mobility models, as explained further.

3.6.1 *Data Sources*

The main data provider for this work is the telecommunication operator, which is going to provide mobile data, which is analysed in Chapter 4. However, in order to characterise the pilot area, a data collection process was taken to gather other more generic data from open-source datasets.

In order to collect the necessary information for the development of this project, open-source datasets were considered. In order to collect the meta-data of the pilot area, one of the main sources is Portal Dados Abertos Administração Pública (2021). This is a portal for free information about several topics in Portugal.

Furthermore, Lisboa Aberta - CML (2021) and Geodados - CML (2021) had significant importance for the characterisation and analysis of the pilot area. The mentioned portals provide public and free information regarding the Lisbon council, which allows spatial study regarding infrastructures, activities, among others. Usually, they return geographic datasets about the city intending to enable the reuse of information produced by the municipality. Several information concerning sports, tourism, health, urban planning, background information and mapping, etc. For that reason, they were the main sources for the data exploration process described in Section 3.6.2.

3.6.2 *GIS Information Analysis*

To analyse several PoI categories in our AoI, it is important to combine them with geographical representation to establish a georeference towards these two parts: the **vectorial infromation** from PoIs, and a **map** that can describe the environment around them.

In this sense, several PoI is going to be studied in this section, such as Education, Mobility and Health, for example. This vectorial information was accessed using the data sources mentioned above, and combined with the map provided by OpenStreetMap (2021), and combined the layers with QGIS software.

Education

Schools and universities can be a crucial location for implementing soft mobility hubs since they are points of traffic during different times along the day and concentrate the typical users, such as young people and students. In the peak hours such as early in the morning or later evening, the number of people engaged on that subject (e.g. teachers, students, school administrators, etc.) is high. Specifically, high school and college, due to the age of people involved, are identified as potential points for traffic congestions and flow.

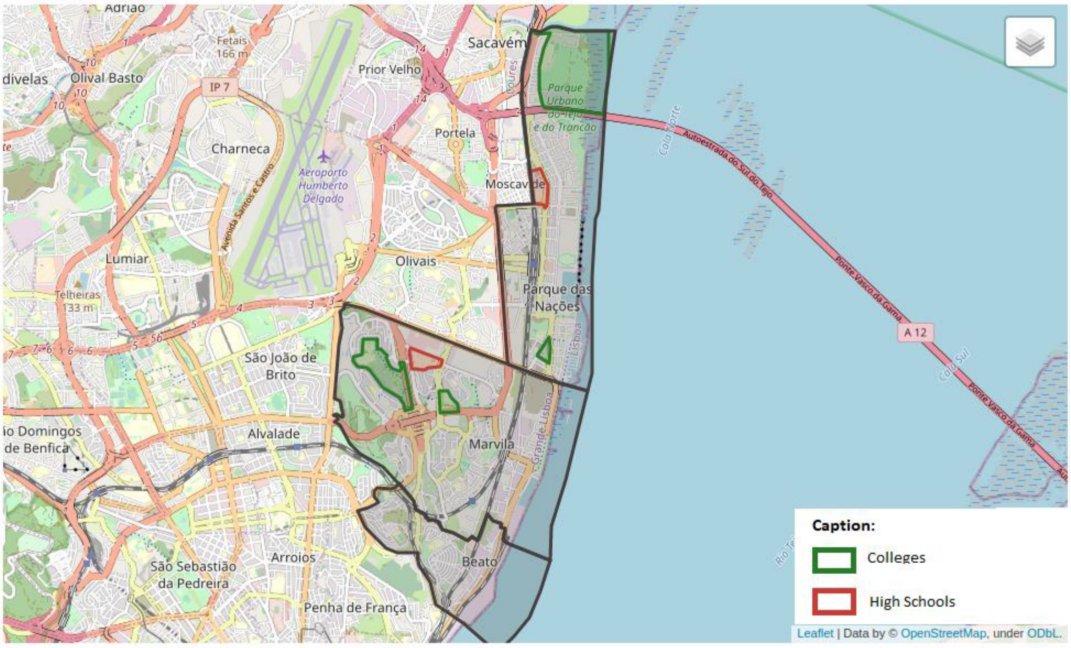


Figure 24: High Schools and Universities in *Beato, Marvila* and *Parque das Nações*.

These places are a key point in society, as we are going to see later in this report. High schools and colleges were taken into account for this analysis since people who move to these places have a more independent way of moving, with a personal car or at least a driver license for it due to their age. Therefore, all the staff related to this type of education is also considerable, since universities and big schools need a considerable number of staff members. In this perspective, primary schools and those who are attended by teenagers were not taken into consideration.

Culture

Cultural spots also represent interesting types of places for picks of mobility demand. As stated before, *Parque das Nações* is a modern and innovative region, where Altice Arena, Lisbon Oceanarium, Camões Theater, among others, are placed. That said, *Parque das Nações* is a major area of cultural interest.

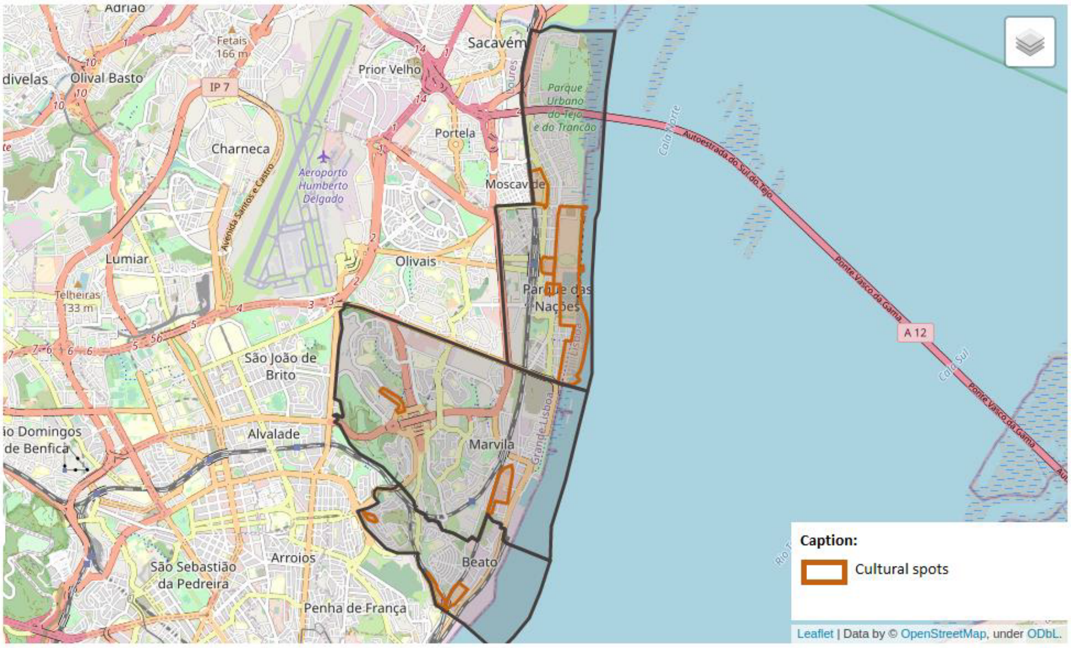


Figure 25: Cultural places in *Beato*, *Marvila* and *Parque das Nações*.

Services

Following the same approach, services can also be attractive for implementing spots for alternative mobility modes and potentially heat points for traffic jams. Supermarkets, offices, moles, train stations, and other services also represent places for high flow, especially during the day. For example, placed in *Parque das Nações*, *Vasco da Gama Shopping* and tram/buses station *Gare do Oriente* concentrates a high number of people.

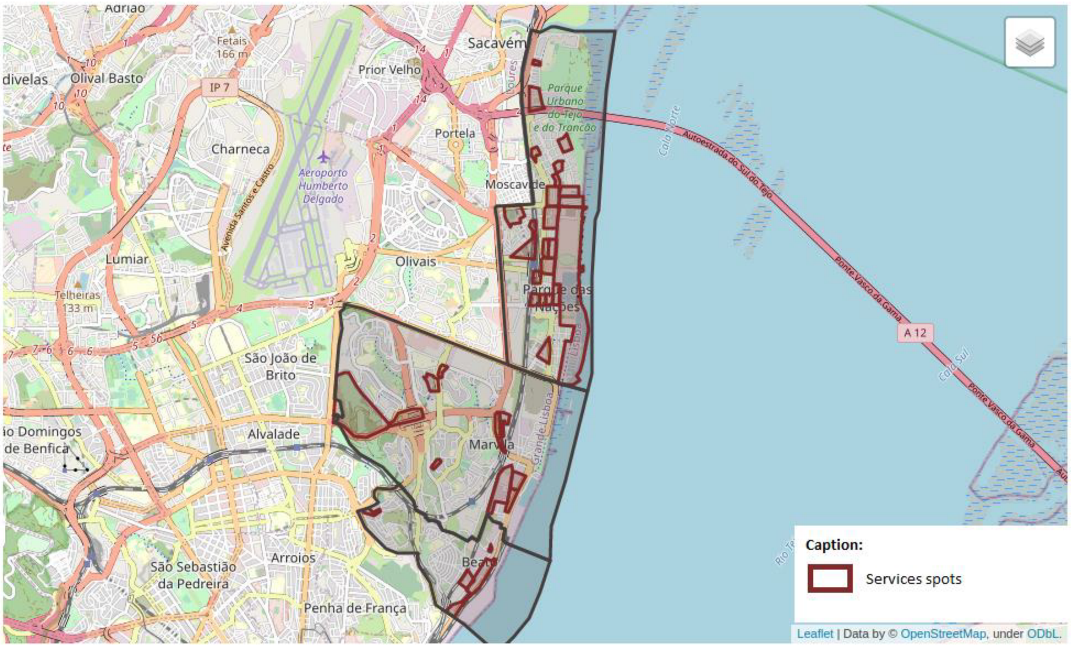


Figure 26: Services places in AoI.

Parking Stations

Parking stations are a very important aspect to consider when studying mobility since soft mobility and other modes of transportation are often seen as intermodal solutions,i.e., as a complement or part of a trip where there is more than one way of transportation. Moreover, it has been proven that the search for parking slots generates a lot of traffic around the city. Therefore, parking bike stations and normal parking stations are present below in order to understand their distribution in the Lisbon pilot area. Recently, EMEL (municipal company in charge of parking, BSS) has launched a pilot to test the acceptance of a monthly fee for surveilled bike parking.

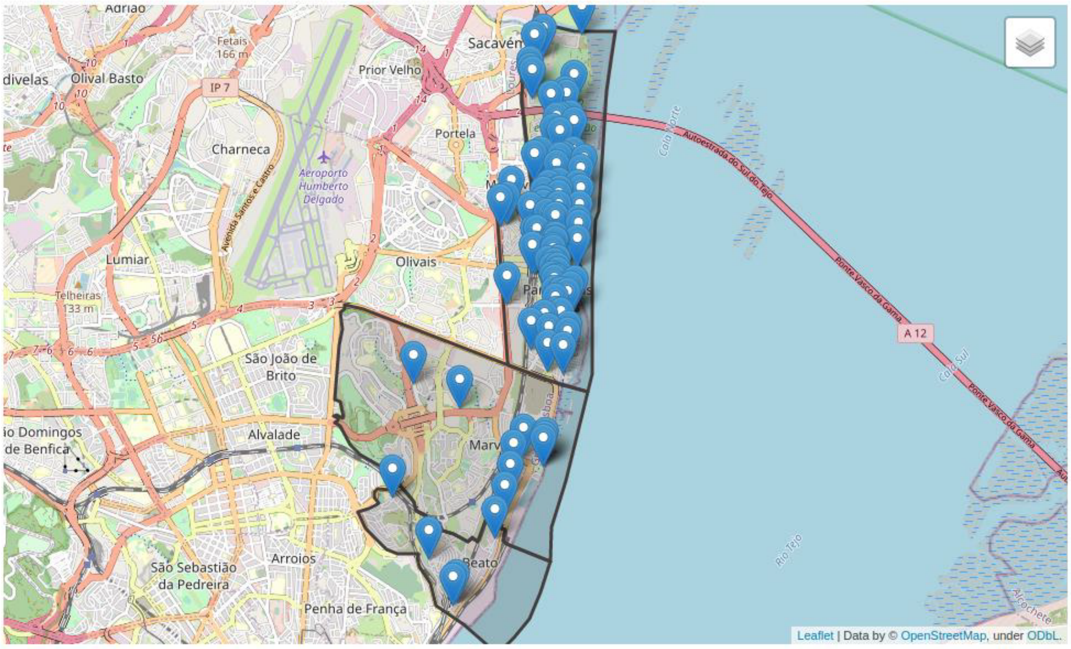


Figure 27: Bike hotspots in AoI.

Around *Parque das Nações*, parking bike hotspots are more concentrated since the favourable slope there and considering the cultural and educational buildings presented previously. *Beato* and *Marvila* show considerably fewer parking hotspots, which can be an interesting way of studying.

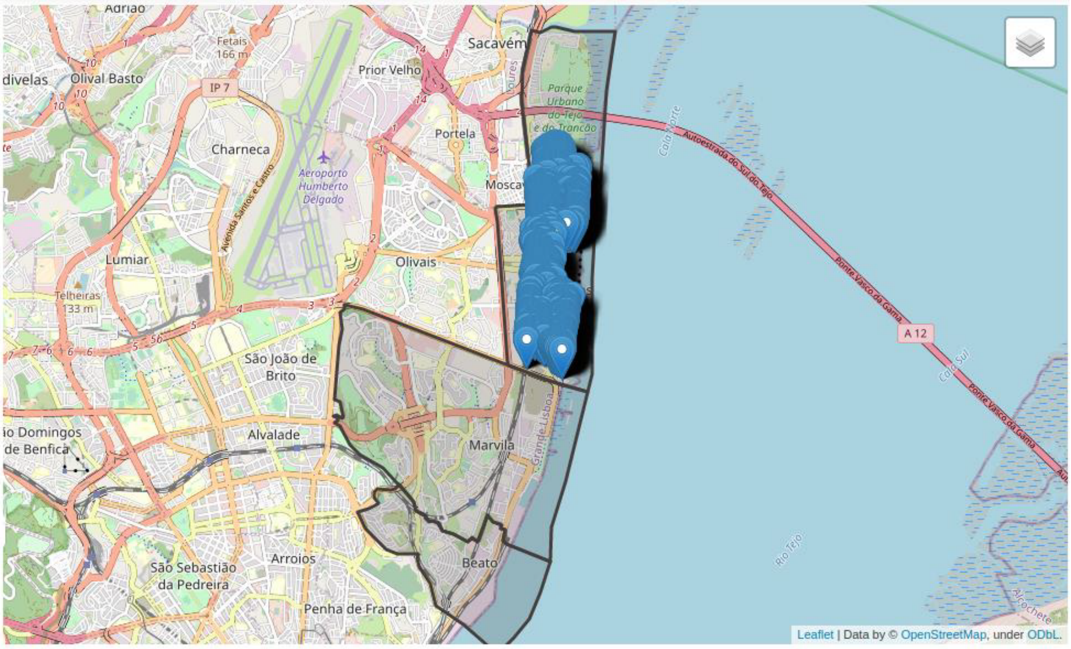


Figure 28: Car Parking in *Beato*, *Marvila* and *Parque das Nações*.

As for parking stations, we can see that all stations that were considered in this baseline take place in *Parque das Nações*. That point should be considered since *Marvila* shows a higher population density, and with the car usage as seen before, parking stations are a good way of handling car parking in this residential zone. Since the commercial and cultural infrastructures placed in *Parque das Nações*, as *Vasco da Gama* shopping, parking stations distribution are much more oriented to this area.

Health

Hospitals and health centres create larges fluxes of people, both staff and patients. Moreover, around hospitals and other health infrastructure usually exist a network of related services such as hotels, restaurants, pharmacies etc., that might increase the fluxes of people moving to these areas.

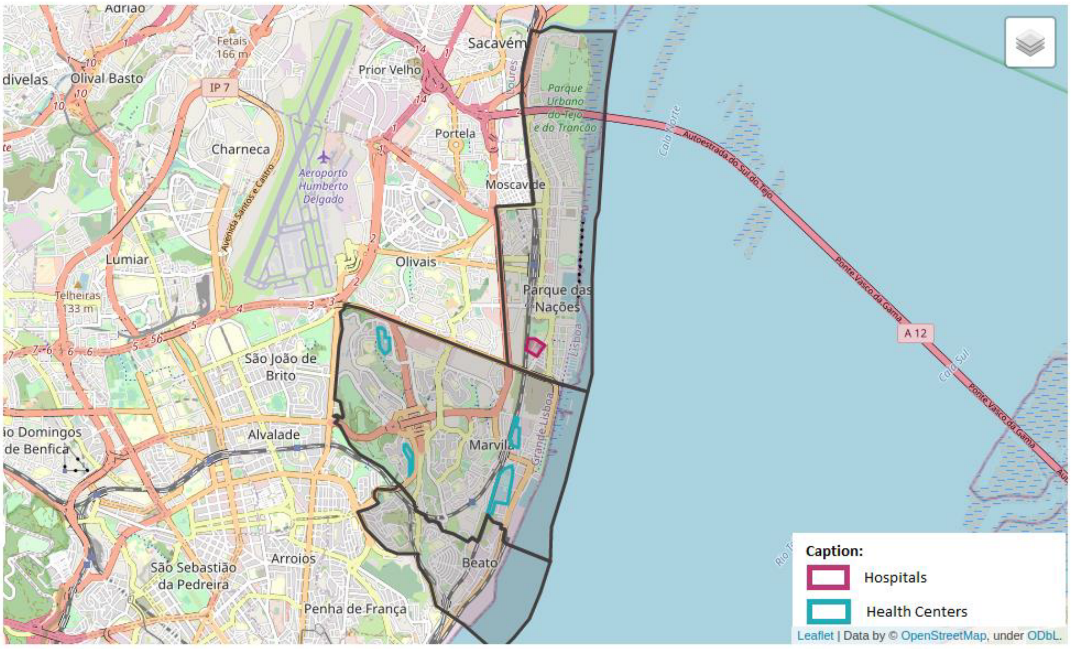


Figure 29: Hospitals and Health Centers in *Beato*, *Marvila* and *Parque das Nações*.

Bicycle Lanes

Apart from public inter-municipal passenger transports' services, Lisbon has also been developing other strategies to develop sustainable mobility. In particular, soft mobility infrastructures have been implemented in the city in order to promote an eco-friendlier mindset. For these reasons and knowing the investment in bike lanes spurs additional cycling, increasing visibility and further reducing risk for all cyclists. Cities like Chicago, New York, Montreal and Paris had started building protected lanes since the beginning of their network, so when they launched their bike-sharing systems, the accession was immediate because people already had safe places to bike.

This means that if a municipality wishes to increase its cycling population, it can invest in a large bike-share system combined with investments in bike infrastructure. Many studies, such as one from the National Association of City Transportation Officials (NACTO (2016)) , point that this equation results in large decreases of risk of injury by cyclists. Providing safety is one of the key aspects to attract more people to adopt new habits of transportation. Given the importance of a substantial bike sharing system, it is possible to transform cycling from a leisure activity to an actual transportation mode. In order to understand our current infrastructures regarding bicycle paths, these are plotted in the Figure below.

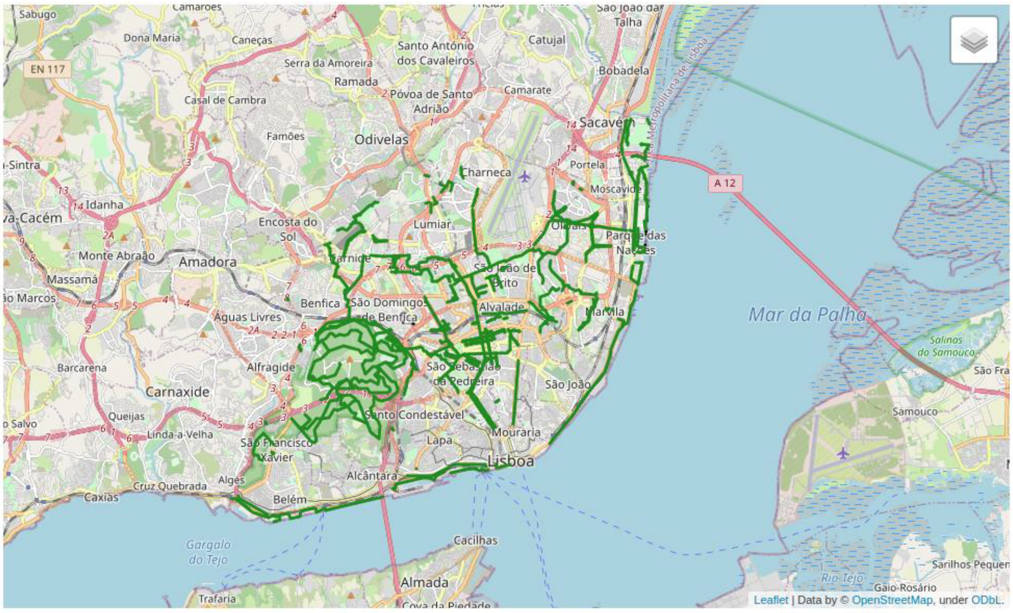
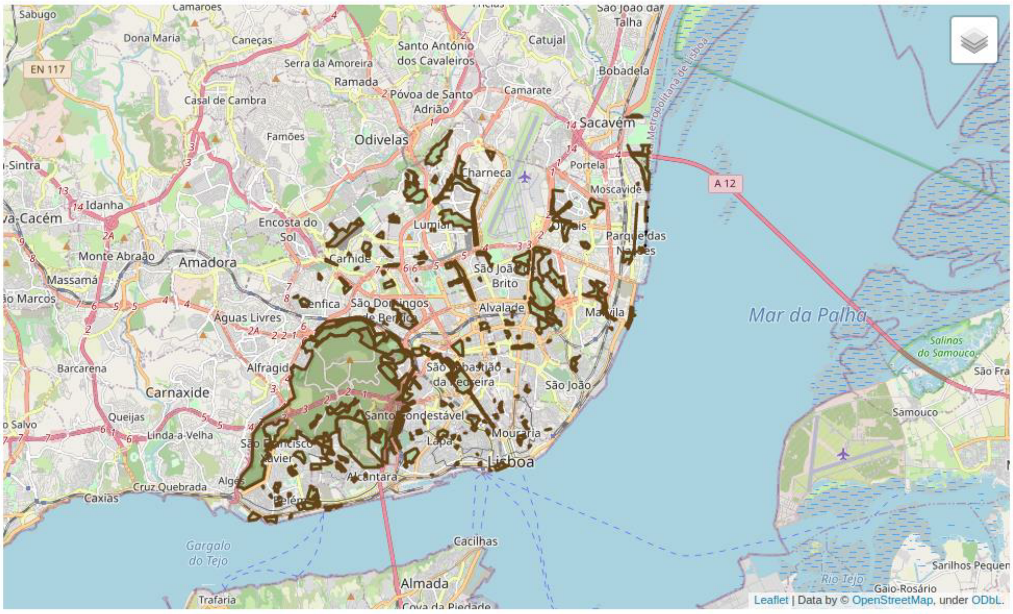


Figure 30: Bicycle Lanes built in *Lisbon* council

This data was collected from Portal Dados Abertos Administração Pública (2021), to briefly analyse the distribution of bicycle lanes in Lisbon. As can be seen from Figure 30, Lisbon has a considerably well-distributed bicycle lane network, and it is expected to grow up to 200km by the end of the year 2021, according to Observador (2021b). That said, it is important to focus on two points. The first one is about the region surrounding the river. Across the north bark, from *Sacavém* to *Algés*, there is a long way to ride, especially located near the river. The second point is the green zone presented on the map, the *Monsanto Natural Park*. As presented, we can see a long bicycle path that is located inside this park, being another initiative to bring people closer to nature.

Green Spaces



Leaflet | Data by © OpenStreetMap, under ODbL.

Figure 31: Parks, gardens and green spaces in *Lisbon* council

From Figure 31, it is possible to verify that Lisbon has several green spaces to walk, practice sports or just be in touch with nature. Regarding the pilot area considered for this study, *Parque das Nações* and *Marvila* have more parks, gardens and green spaces in general. In the *Beato*, there is only one green space represented. From that information, we can perspective a more leisure area in *Marvila* and *Parque das Nações*, since we can combine green spaces and bicycle paths data, being that two areas the most favourable areas to consider.

4

DATA EXPLORATION

4.1 DATA STRUCTURE

From the Portuguese telecommunication operator, a dump from real mobile data was provided, and it was at the raw state. However, at the initial phase, it was possible to verify all the available metrics and parameters that we were able to be used. In Table 2, we can find all the parameters available for this study and their meaning in order to easily understand which constraints and limitations are going to be faced.

Table 2: Data structure and meaning of each attribute

Column Name	Data type	Meaning
Identifier	Integer	Unique identifier for each record in dataset. It has no correlation with devices identification.
Day	Date	Date when device was detected in the given area.
Hour	Integer	Corresponds for all the available time frames during the day. Temporal granularity for all records.
Speed	Float	Means the average speed of the devices detected in a given area.
Number of Devices	Integer	Indicates the number of aggregated devices for specified entry in the dataset.
Bin	MultiPoint	S2 Cell defined by list of points given in ESPG:4326. Defines an area similar to a square of 10x10m.

For a better understanding of each data type and column organisation from the GSM data provided, Table 3 illustrates some sample.

Table 3: GSM data illustration: column organisation and types

Id	Day	Hour	Speed	Number of Devices	Bin Geometry
4324234	26/01/2020	12	5,67	3	MultiPoint ((-9.12,38.75), (-9.122,38,7561),...)
4324278	27/01/2021	00	0	10	MultiPoint ((-9.11,38.45), (-9.111,38,451),...)
4324218	27/01/2021	15	33,06	5	MultiPoint ((-9.01,38.05), (-9.011,38,051),...)

4.2 VISUAL ANALYSIS

4.2.1 *Speed Distribution Analysis*

As mentioned previously, one of the important factors to take into account is speed. That said, and according to the several data we have access to, let's focus on "mean speed". This metric represents the average value for speed registered on each entry of the dataset, meaning (if it was measured correctly) the average speed that the person or the group of people registered when it was monitored.

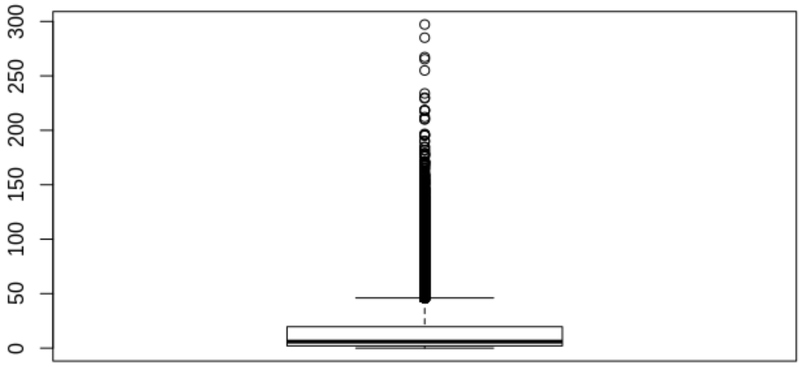


Figure 32: Mobile data mean speed distribution (raw)

As we can see in Figure 32, there is a massive range of data, according to the boxplot presented above. The major amount of data corresponds to the interval of 2 km/h to 20 km/h (quantile 25 and 75, respectively). However, we observe a much larger distribution of points after quantile 75, being from 30/40 km/h (quantile 95 is 47 km/h) until 300 km/h. This means that, in order to use data properly, it will suffer a clean and preprocessing process because it is probably misinformation and corresponds to outliers and is even incorrectly measured. Using certain and clean data is very important since we have not any information about the vehicle that people were using if they were riding a bike or just walking. For that reason, the use of speed is a very important topic in this project regarding the lack of other information. In this sense, we must be careful when using the speed variable and its values.

4.3 GEOGRAPHICAL DISTRIBUTION

4.3.1 *First Insights*

As we can see in the Figure 33, we have a large distribution of points in our pilot area (*Parque das Nações*, *Marvila* and *Beato*), but also in the surrounding parishes. The provided data allows us to understand the coverage geographically, being possible to study mobility in our study area. Moreover, we could also study the surrounding areas and possibly the variation of flow during the days due to the large spatial distribution of provided mobile data.

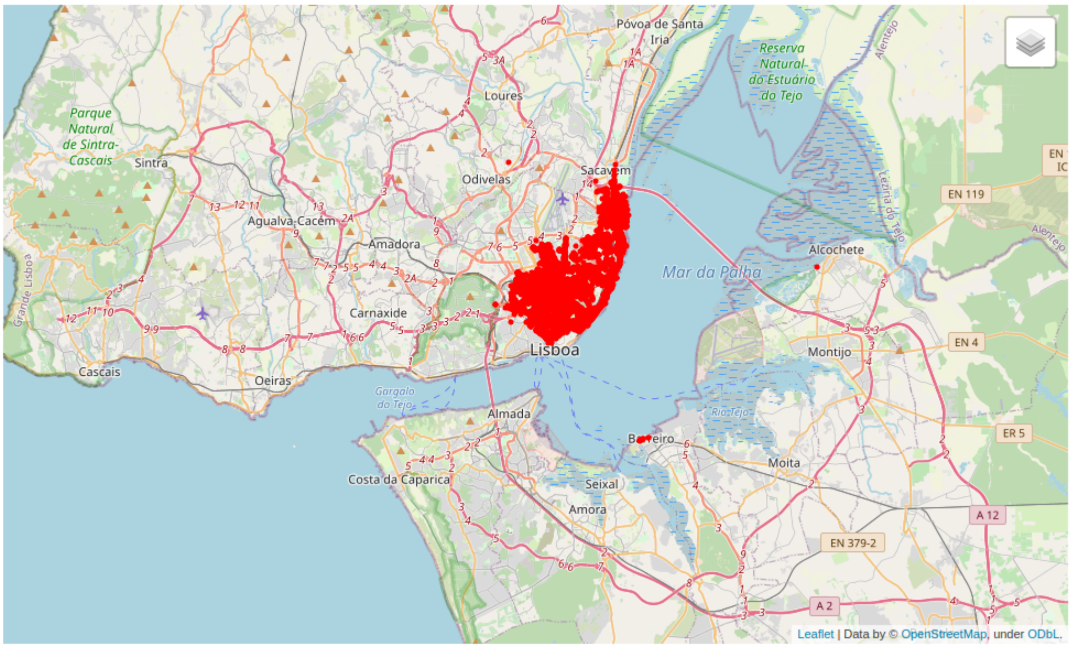


Figure 33: Mobile data distribution at midnight in LMA.

4.3.2 *Number of Devices Distribution*

As stated before, one of the available data is "number_of_devices" that were considered in a determined area at the determined time. This technique of aggregation has been done by the Portuguese telecommunication operator, following a concept named **bin**, which represents the geographical geometry of each record, and it is very similar to a square, and defines an area where the number of devices was monitored. This is an imposed limitation due to privacy and GDPR in place in Europe. That said, it was possible to draw the distribution according to the number of devices. For this example, different colours were taken into consideration, being the colder ones related to lower values of people concentrated, whereas the colder ones regarding the higher number of people and consequently more affluence area.

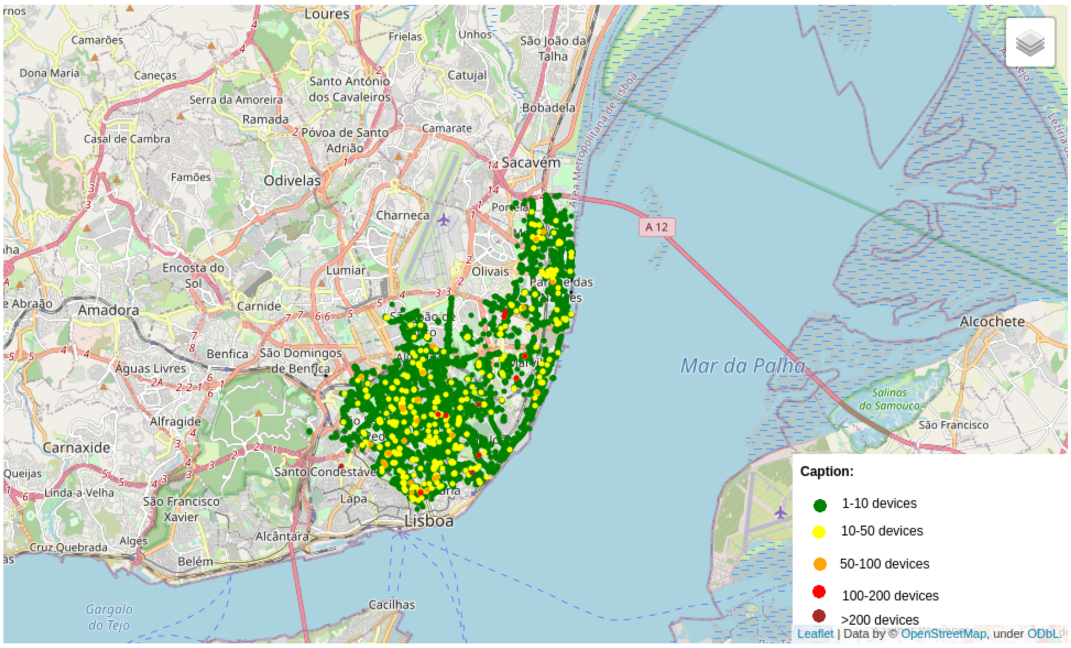


Figure 34: Data distribution in pilot area and surroundings at midnight and number of devices aggregated.

As we can see in Figure above, hotter colours represent a higher number of devices. There are a few points with red and brown and a lot of them with green and yellow. For this reason, it is difficult to understand the real distribution of people since several green points would represent a yellow one, making analysis harder to figure out. For that reason, another approach was taken into consideration: use a single colour for all points, but change opacity according to the number of devices aggregated. This means lighter areas are going to represent fewer people concentrated, whereas darker areas mean more concentration of people. In order to have a comparable scenario, two different hours of the day are presented below, being the left one regarding midnight and the right one at noon.

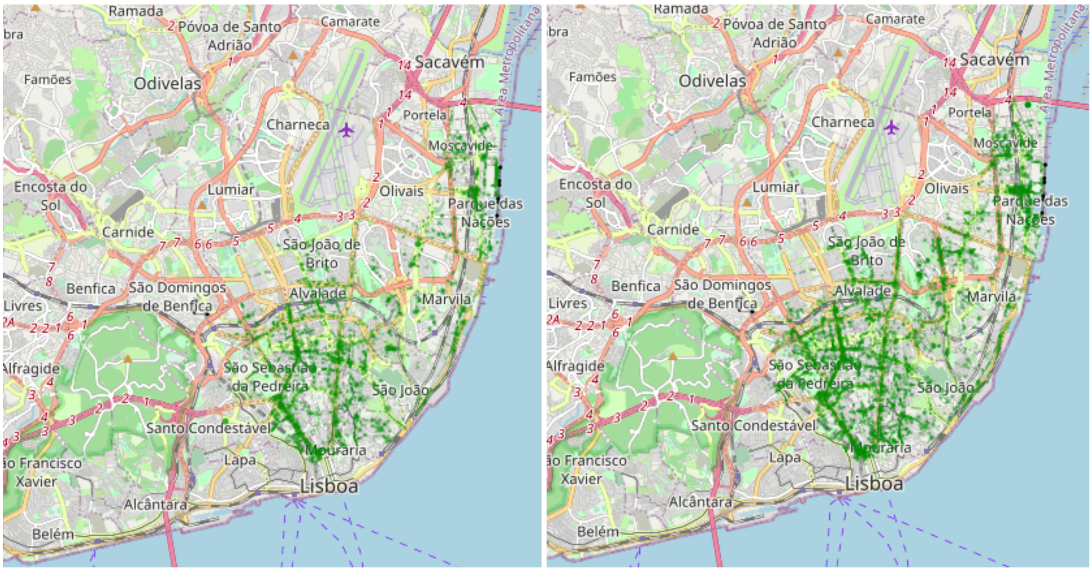


Figure 35: Devices distribution comparison: 0 am vs 12 pm .

As we can see in Figure 35, there are several differences between these two time frames, in the same day. On the left, we have the supposed distribution where people are on their houses, and of course where the flow outside is really low. On the right, we have probably one of the most representative hours of people working on their job places and being of course, outside their houses.

For that reason, we can define some interesting areas in the right Figure, being them close to *Mouraria*, with a significant darker area, but also in our pilot area, *Parque das Nações* represents a high affluence in the noon time.

That can be corroborated by the total amount of companies in Lisbon, according to *Câmara Municipal de Lisboa* (2020), which is illustrated in Figure 36.

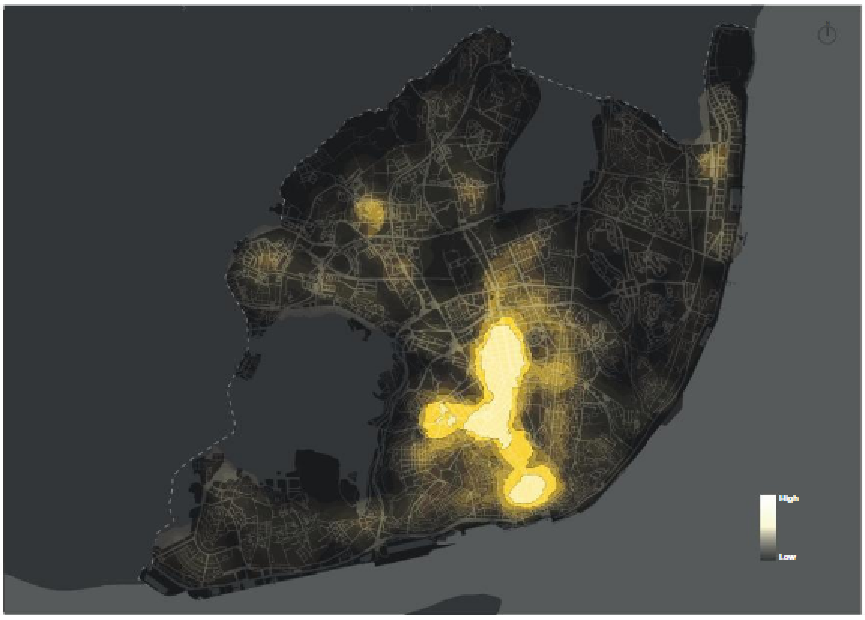


Figure 36: Companies distribution in *Lisbon* in 2019.

Source: *Câmara Municipal de Lisboa* (2020)

4.4 COUNCIL PARISHES INDIVIDUAL ANALYSIS

This section aims to illustrate some interesting areas in terms of mobility. This analysis establishes some correlations between the spatial distribution of devices, according to their area and the time of the day. For briefly comparisons, four different hours are going to be considered: during the night (2 am), in the morning (10 am), in the afternoon by 3 pm and also at 8 pm. These periods were selected in order to capture the several mobility patterns available during the day and night in our study area in order to identify main roadways, residential and business areas. As known, the Area of Interest (AoI) for this thesis consists of three parishes in oriental Lisbon city centre: *Beato*, *Marvila* and *Parque das Nações*.

4.4.1 *Beato*

The first parish to be analysed is *Beato*. This is a small parish in Lisbon, the smallest one among the three belonging to AoI, as presented in Section 3.5. As we are going to see, we can see some interesting differences between mobility patterns, according to the time frame considered, as should be expected.

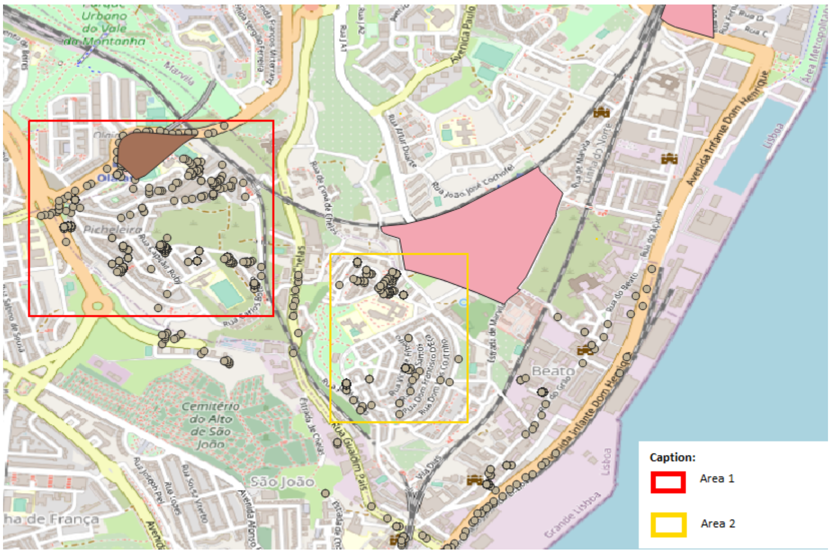


Figure 37: Beato 2 am: Interesting areas identification

As we can see in Figure 37, we have two different areas of interest. The first one, Area 1, which is coloured with a red rectangle, represents a considerable high concentration of devices during the night, representing a potential residential area, which is more populated during the night. For that reason, it is also important to consider the closeness of a major part of these devices to the Olaias' Metro Station, which is coloured brown. Oh the other hand, we have chosen a second interesting area of study, Area 2, to place our sensors, which is close to Marvila's Train Station (yellow rectangle). This is an interesting area during the night, with a considerable concentration of NOS devices, but also during the day, as we can see in Figure 38.

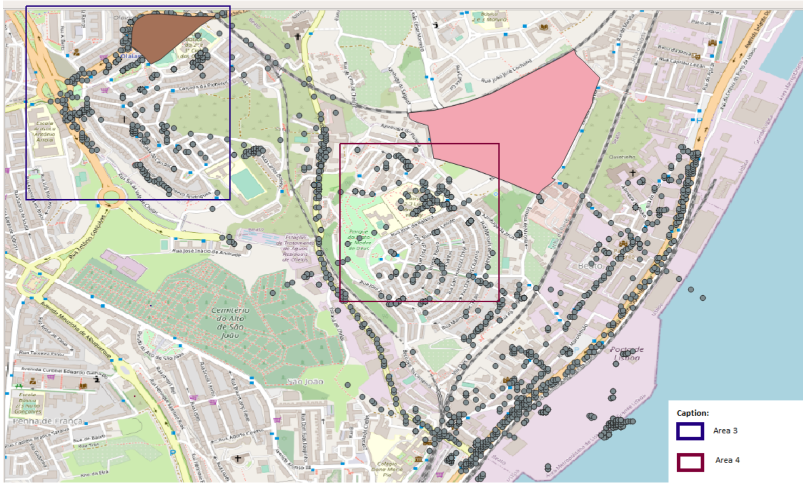


Figure 38: Beato 10 am: Interesting areas identification.

As we can see in Figure 38, there are plenty more points at 10 am. For this time frame, during the morning when it is expected most of the people are on study/workplaces, we can identify two areas of significant agglomeration of people. Area 3 is marked as a dark blue rectangle, which represents the surroundings of the metro station (on brown). The second area, Area 4, is much similar to the one identified at night (Figure 37), which means that this is gathering a considerable high number of people during the night, but also during the day, because of the residential infrastructures near the school, in centre of the dark red rectangle. One more important factor to consider is the main gateways and roads in *Beato*.

As we can see, there are three important streets in *Beato* that represents a huge part of the flow in this parish. On the left, near the dark blue rectangle, we have *Avenida Engenheiro Arantes e Oliveira*, crossing the middle of *Beato*, we have *Estrada de Chelas*, and near the river, we have *Avenida Infante Dom Henrique*, which is a very important way to connect *Beato* and *Marvila*, as we are going to see further, but also an important road because it has alongside its extension and industrial area with considerable business and services, and due to closeness to river and *Porto de Lisboa*.

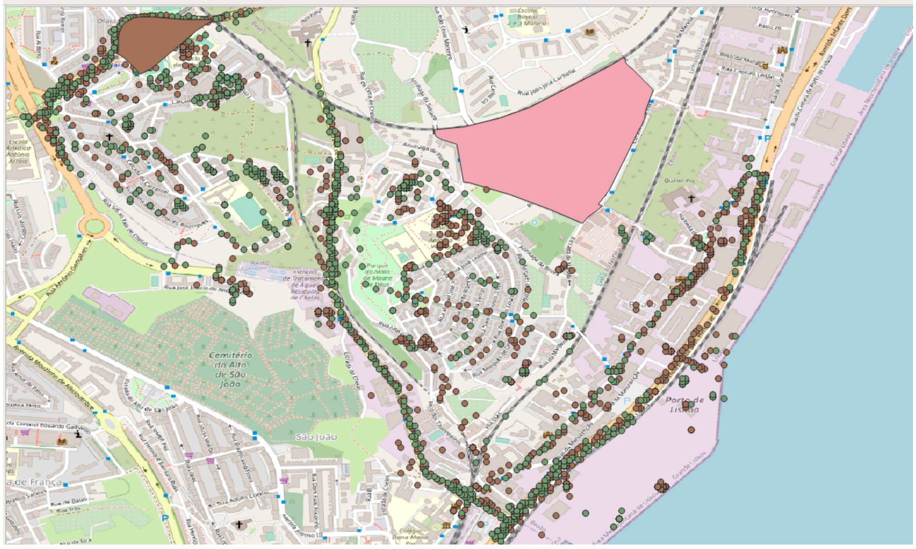


Figure 39: *Beato* devices' distribution at 3 pm (brown) and 8 pm (green)

As shown in Figure 38, we have a comparison between 3 pm and 8 pm, which are two important time frames during the day. Firstly, we can have a global interpretation on localisation of people studying/working (similar to 10 am), but also a direct comparison with a late hour, where it is normal people returning to their houses. As presented before, we have an interesting concentration close to the metro station, both at 3 pm and also at 8 pm. On the other hand, close to the school*Escola Básica 2-3 Luís António Verney*, which was previously identified as a possible interesting focusing study. Also, we can verify high flow in the main roads in *Beato*, mainly focusing in the *Estrada de Chelas*, crossing the parish on its centre, but also *Avenida Infante Dom Henrique*, close to the river and having a high number of people using it, since it is a considerable fast way of moving in *Beato*.

4.4.2 *Marvila*

Marvila is our second parish which is undergoing analysis. This is a bigger area when compared to *Beato* and also more populated, according to official information provided by INE (2011). Firstly, we are going to focus as presented before: night distribution of NOS devices, following the morning, afternoon and evening.

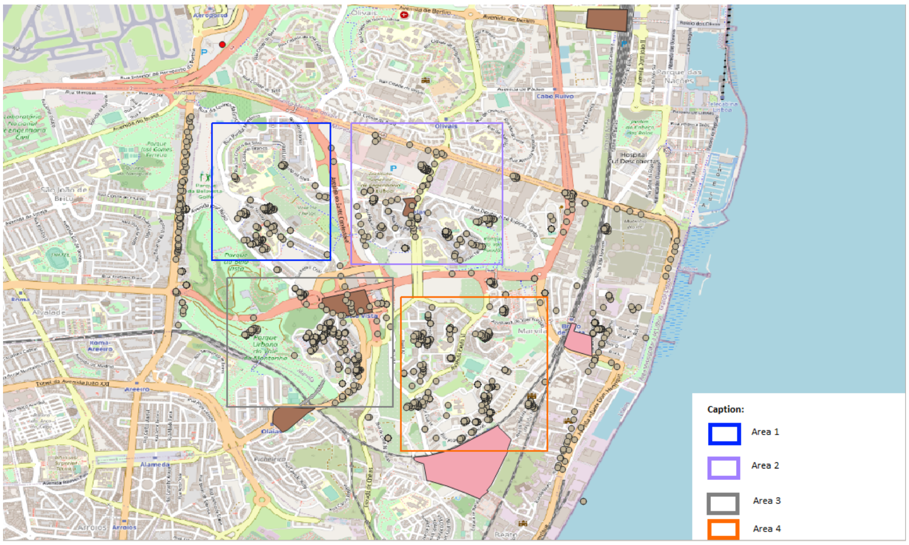


Figure 40: *Marvila* devices' distribution at 2 am .

As presented in Figure 40, we can identify four potential areas of attractiveness in terms of habitability and residential spots. As we can see, in Area 1, it is possible to identify a concentration point, which corresponds to a social house building, *Bairro da Flamenga*. In Area 3, the amount of people is possibly due to closeness to metro stations, and also the fact of having an elementary school and kindergarten, being them *Escola Básica 1 Bairro do Amador* and *Jardim de infância Bairro do Amador*. As these names suggest, these buildings are located in *Bairro do Amador*. On the other hand, Area 2 and Area 4 also represents proximity to metro or train stations, but also another interesting infrastructure that can be taken into account. Specifically, in Area 4, we can find two points of agglomeration, which are very close to natural spaces, in fact, playgrounds, representing the existence of houses nearby these spaces. In Area 2, we can find interesting locations regarding education, but also an engaging fact: a considerable number of devices detect on *Marvila*, a fast food shop. Being objective, it is McDonald's *Marechal Gomes da Costa*. Considering the time frame presented, it may be due to drive-thru service. As we have already stated before, the orange line on the right representing *Avenida Infante Dom Henrique* still is an important roadway in this area, linking *Beato* and *Marvila* nearby and alongside the Tagus river.

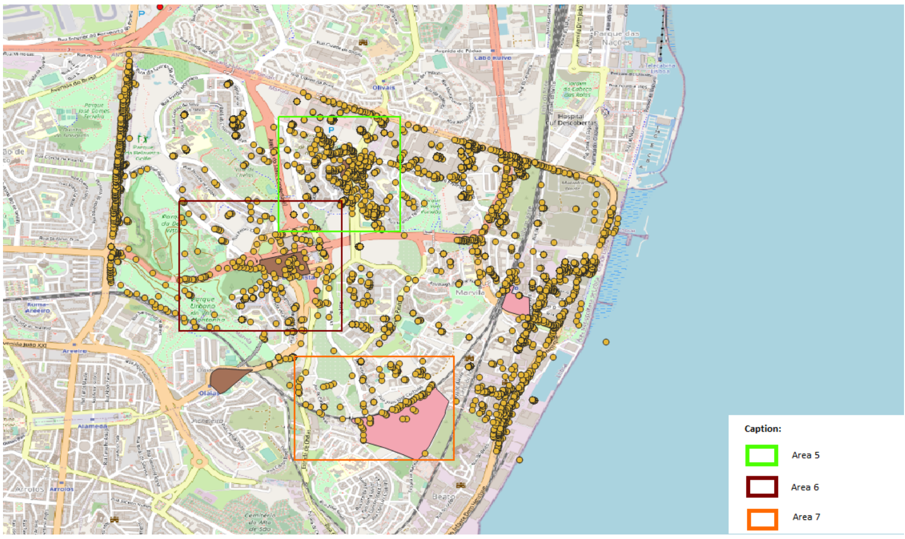


Figure 41: *Marvila* devices' distribution at 10 am .

When comparing night to morning, we can identify several differences and affirm a noticeable contrast between these two-time frames. During the day, specifically in the morning, we observe much more devices density in general, with a particular focus on some areas. As shown in Area 5, there is a massive concentration surrounding the northeast metro station in *Marvila*, *Chelas*. This phenomenon occurs due to the presence of a metro station but also a university location nearby the metro station. Area 6 aggregates Bela Vista metro station and also the evolving area, whereas Area 7 is identified with *Marvila* train station. It is important to notice the difference in traffic congestion in the identified roadways, such as *Avenida Infante Dom Henrique*, but also the other major access, which can be *Avenida Marchal Gomes da Costa* and *Avenida Almirante Gago Coutinho*. These are very important roads because they circumscribe *Marvila*'s parish.

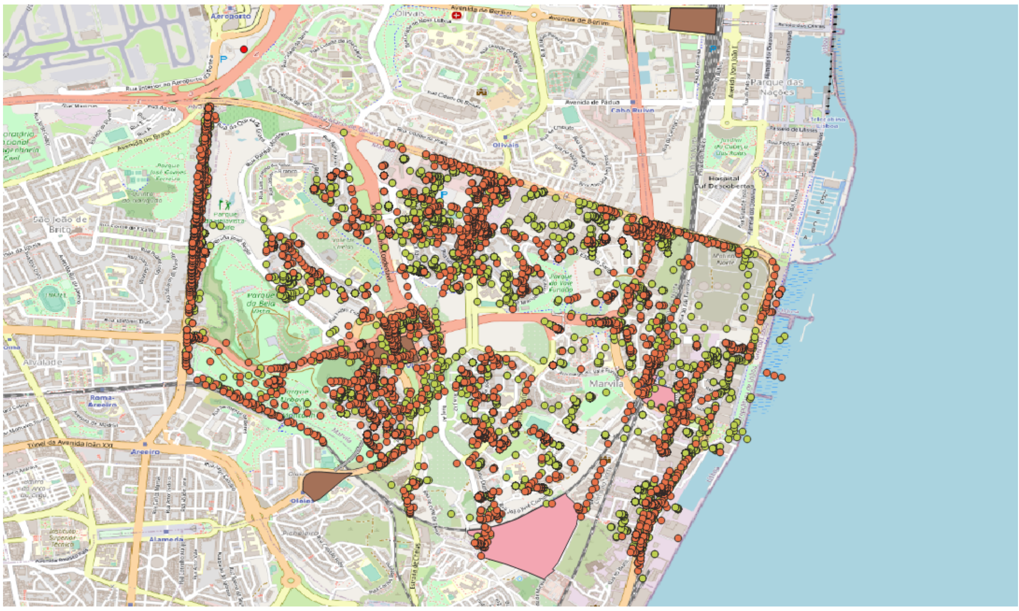


Figure 42: *Marvila* devices' distribution at 3 pm (green) vs 8 pm (orange) .

In Figure 42, we present a comparison between the middle afternoon hour and evening hour. As we can see, the regions close to Metro stations in *Marvila*'s centre is high, at both hours. On the other hand, the main access of the parish is also very populated on both time frames.

4.4.3 *Parque das Nações*

This is our final parish in the AoI. *Parque das Nações* is a very particular area in Lisbon, being one of the most iconic places in Lisbon city centre, due to its commercial and cultural environment.

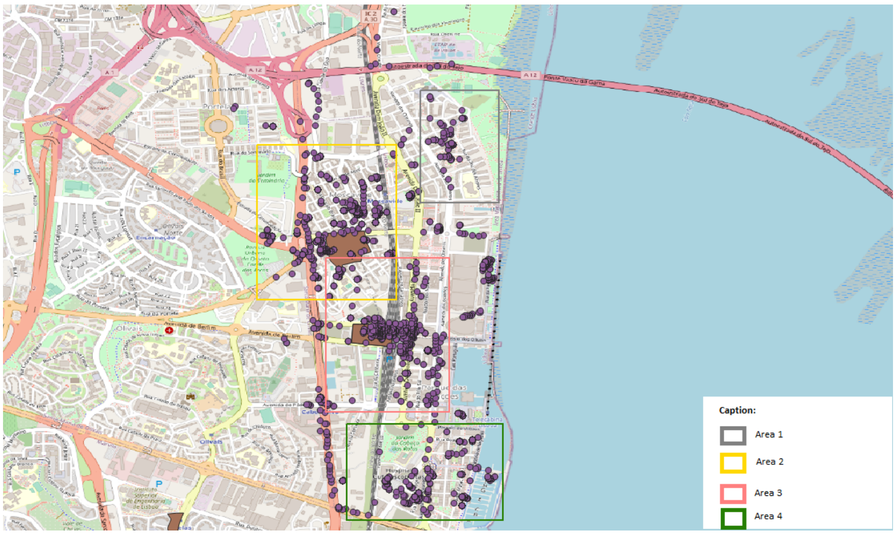


Figure 43: *Parque das Nações* devices' distribution at 2 am .

As we can see in Figure 43 presented above, it can be identified four different areas, considering different locations and surroundings. Firstly, Area 1 represents a residential area with some buildings. In Area 2 (yellow), we can recognise high concentrations nearby Moscavide's Metro Station. Also, it is important to highlight one of the biggest arterial roads in *Parque das Nações*, *Avenida Infante Dom Henrique*, which goes until *Marvila* and *Beato*, as we saw previously. Area 3 (pink) is also very major in *Parque das Nações*, due to the number of people inside this area, but also due to its closeness to G*are do Oriente* - train and metro station, but also *Vasco da Gama Shopping* and other cultural infrastructures. In the middle of *Parque das Nações*, parallel with *Aevnida Infante Dom Henrique*, we have *Avenida Dom João II*, another important road in this area. Area 4 (green), by its turn, represents closeness to other services and health spots in this parish. It is close to the river and *Hospital das Descobertas*, and residential buildings too. *Parque das Nações* is the most attended parish of our three. Due to its spaces buildings, business infrastructures, cultural buildings and commercial development, it represents a considerable part of traffic flow in Lisbon. From night to day, it shows an incredible difference in terms of devices detected.

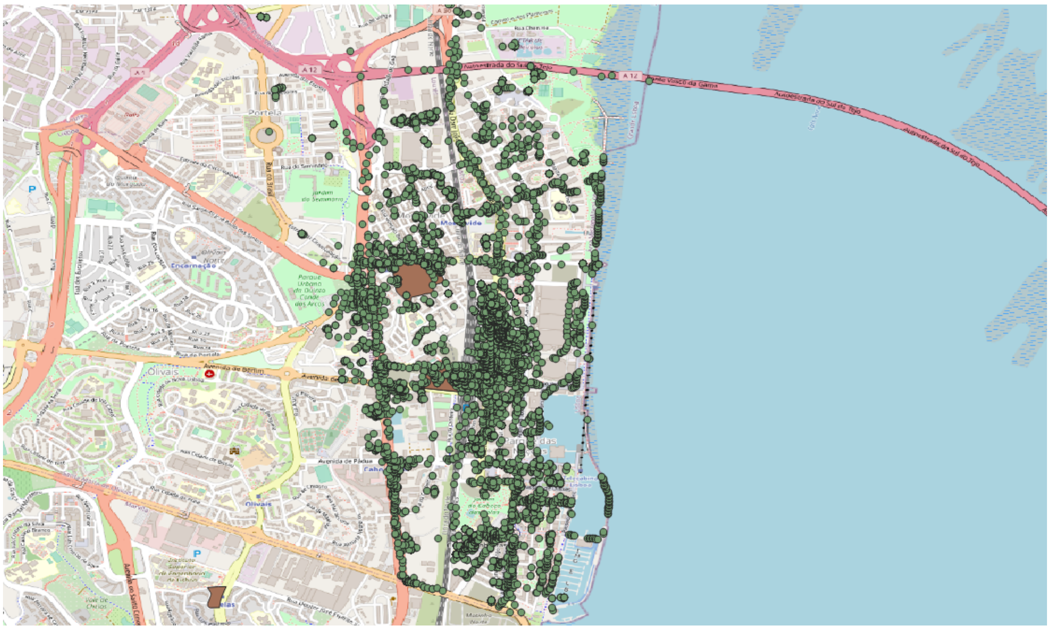


Figure 44: *Parque das Nações* devices' distribution at 10 am .

As shown in Figure 44, there are a lot more devices plotted. The main roadways are now much more painted, being almost covered by coloured dots, representing devices. The avenues mentioned previously are much more attended by their time frame, and the areas identified are all much more painted. Despite being hard to identify factual areas with more attendance, it is clear that areas nearby metro stations (brown) attract a lot more people.

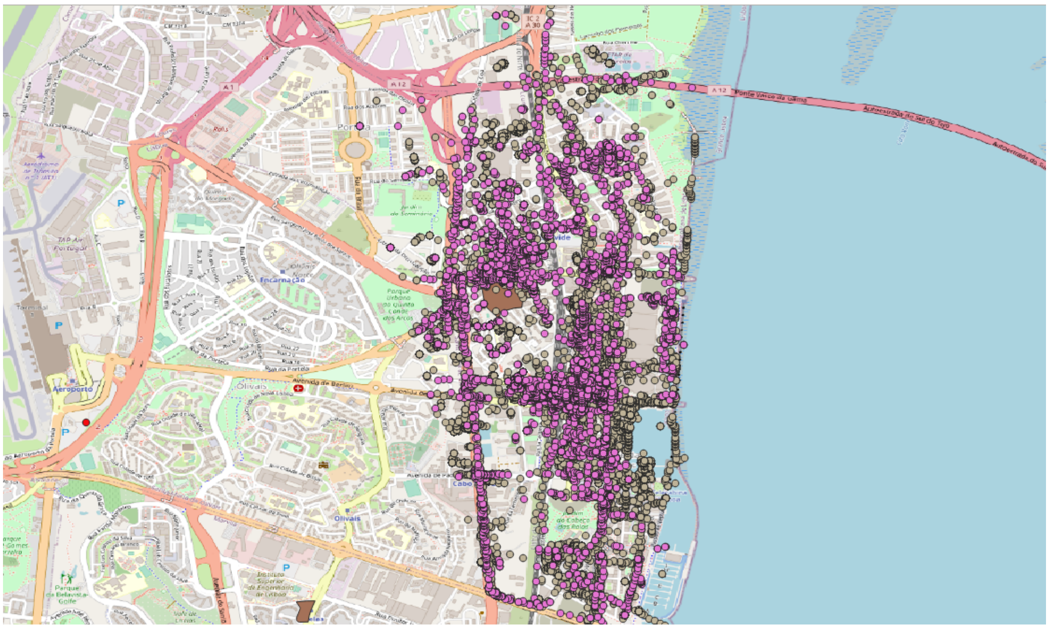


Figure 45: *Parque das Nações* devices' distribution at 3 pm (grey) vs 8 pm (pink) .

In Figure 45, we can still see a massive distribution at 3 pm (grey) and also at 8 pm (pink). This means *Parque das Nações* presents a much higher flow during more time when compared to our two other parishes. At dinner time, we can even spot a high number of devices outside residential areas, as shown in Figure 6, which proves the hypothesis that these devices still be in a move at that time.

4.4.4 *Summary of Main Areas Interest*

In order to summarise the main areas identified during the visual analysis of GSM data regarding devices concentration during day and night time frames, Table 4 provides an objective description of each area, taking into consideration major nearby PoI and giving a brief description for each identified area, justifying their importance and their location.

As Table 4 indicates, at least three major areas have been identified in the visual analysis process for each council parish, showing a better correlation between the PoI and the concentration of the devices for a given time. Also, an important look at the existing roadways and bicycle lanes nearby these areas were considered so that it was possible to understand the potential for each area in terms of soft mobility solutions installation in the future and also the present.

Id	Parish	Coordinates/Area	Characterization	Main target
1	Beato	Olaia's Metro Station Escola Básica 2-3 Olaias	- Close to School - Close to Metro Station - Major people flow - Close to important roadway (Avenida Engenheiro Arantes e Oliveira)	People
2	Beato	Escola Básica Luís António Verney and close residential buildings	- Residential area - Reasonable amount of people during night and higher during day - Nearby Train station	People
3	Beato	Avenida Infante Dom Henrique	- High traffic flow during day - Close to river - "High-speed" avenue - Closeness to industrial and business area - Access to Porto de Lisboa - Major cars flow	Vehicle and Bicycles
4	Marvila	Bairro da Flamenga Escola Básica Luísa Neto Jorge	- Residential Area - Close to kindergarten and elementary school - High number of people during night	People
5	Marvila	Chelas Metro station Instituto Superior Engenharia Lisboa	- Close to university - Higher number of people during day - Fast food shop attraction: day and night	People
6	Marvila	Bairro do Armador	- Residential Area - Close to elementary school and kindergarten - Higher number during day - Close to Parque Urbano Vale da Montanha	People
7	Marvila	Avenida Marchal Gomes da Costa	- Major car flow - Higher traffic flow during day	Vehicles and bicycles
8	Parque das Nações	Moscavide's Metro Station	- Close to Metro Station - More residential area	People
9	Parque das Nações	Gare do Oriente Vasco da Gama Shopping	- Cultural and business spot - Close to train, metro and bus station - Considerable high flux during night - Super high flux during day - Relatively close to other cultural interesting spots	People
10	Parque das Nações	Avenida Dom João ll	- Central roadway - High number of buildings on both sides - Major car traffic flow - Multiple traffic lanes	Vehicle and bicycles

Table 4: Summary of main areas characteristics in AoI regarding devices concentration

5

METHODOLOGY

5.1 METHODOLOGY OUTLINE

As stated previously, mobile data was received by a Portuguese telecommunication operator. The provided data contained more than 700 000 samples, being relative to January of 2020 for almost the entire LMA. However, due to privacy policies and GDPR procedures, this **dataset** is completely **private** and **cannot** be **accessible** by any other authorised member of the project in which this dissertation is inserted.

In this context, the first step consisted of developing a methodology pipeline, which started with acquiring a data-gathering mechanism, then the data was cleaned and analysed.

Some parameters were kept to be used during the analysis and cleaning process, and others were removed (e.g. time frame availability of the data, external factor, public holiday, etc., since they would affect mobility patterns).

The provided data have two distinct components, as already mentioned in Chapter 4, being them:

1. **Vectorial Data** - represents the **shapefiles** and other vectorial files that contains the geographical representation of our data.
2. **Non Vectorial Data (Tabular)** - includes all the available data that is represented in a tabular structure, which includes "Day", "Hour", "Mean Speed", "Number of aggregated devices", etc.

Due to privacy policies and according to GDPR, it was not possible to work with exact geographic coordinates. Therefore, it was necessary to consider each S2 cell, the **geometry** column of our **Non-Vectorial Data,** which is represented in the shapefiles, as the geographical characterisation of each record (area). Figure 46 illustrates the initial representation of the vectorial data.

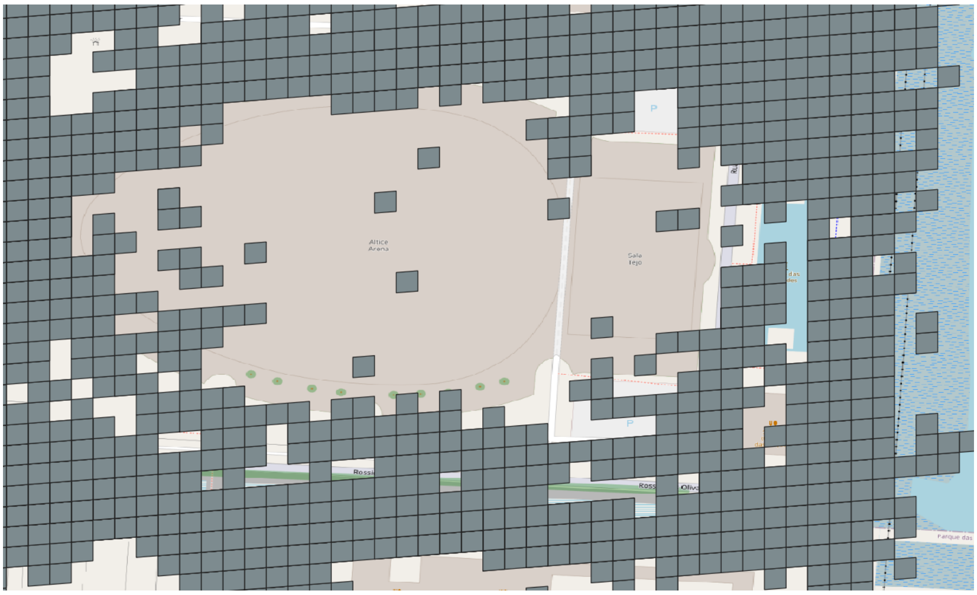


Figure 46: Initial representation of vectorial data: S2 cells (Polygon)

As represented in Figure 46, the geometry used is a Polygon, which defines the boundaries of each record (similar to a square of 10 meters side length).

Using area parameters for each record, it was verified that it added extreme complexity to the analysis since more than one point was being considered for each record, resulting in a MultiPoint structure type. This point led to a necessity for complexity reduction (centroid calculation), which is represented in Figure 47.

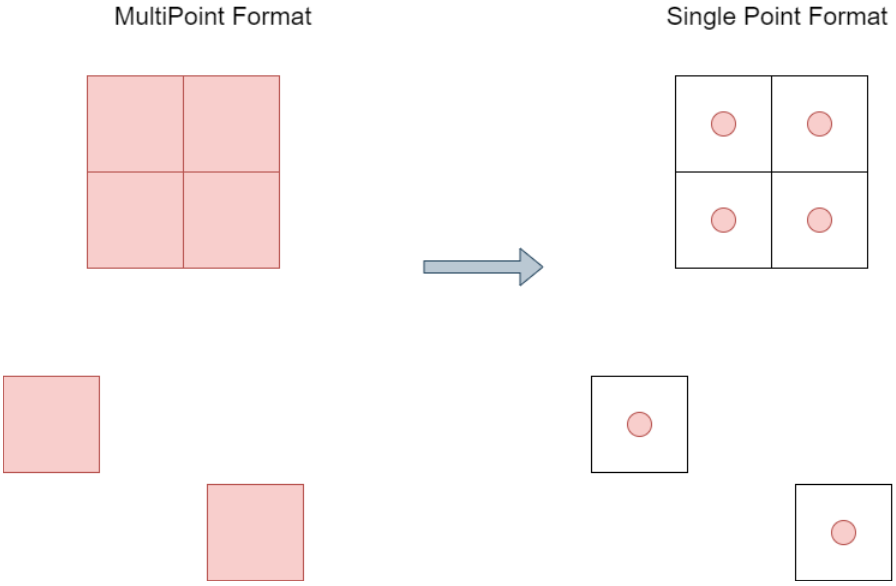


Figure 47: Vectorial Transformation of MultiPoint to Single Point geometry information

The centroid is also known as the "centre of gravity" or the "centre of mass". The position of the centroid assuming the polygon to be made of a material of uniform density, which is the case, is given below. The centroid calculation was performed aided by GIS software (finding the centre of each polygon for sets of individual points), following:

$$C_x = \frac{1}{6A}\sum_{i=0}^{n-1}\left(x_i + x_{i+1}\right)\left(x_i y_{i+1} - x_{i+1} y_i\right)$$

$$C_y = \frac{1}{6A}\sum_{i=0}^{n-1}\left(y_i + y_{i+1}\right)\left(x_i y_{i+1} - x_{i+1} y_i\right)$$

and where A is the polygon's signed area,

$$A = \frac{1}{2}\sum_{i=0}^{n-1}\left(x_i y_{i+1} - x_{i+1} y_i\right)$$

As a result, Table 5 describes the data transformation regarding the geometry simplification process **after centroid calculation**. As possible to verify in the following table, each sample has been converted from a MultiPoint feature, i.e. a polygon, into an **atomic data point**, which is represented by a single point according to the geographical representation. This is a **crucial** step in order to get atoimc data points that can be **easier** used and managed in order to achieve the defined results, since they represent a **downgrade** for the **geographical complexity** of the current problem.

Table 5: Centroid generation process for geometry column

Id	Day	Hour	Speed	Number of Devices	Bin Geometry
4324234	26/01/2020	12	5,67	3	Point ((-9.123,38.755), (-9.1225,38,75615))
4324278	27/01/2021	00	0	10	Point ((-9.11,38.455), (-9.111,38,4515))
4324218	27/01/2021	15	33,06	5	Point ((-9.01,38.055), (-9.011,38,0515))

After the previous point, it was necessary to create an equitable distribution of our device points (since they have been aggregated), then an unlist operation was performed. In this sense, each entry was repeated throughout the dataset N times, being N the number of devices considered in that capture. Table 6 presents the methodology used in this work. The **output** from this operation for the **sample** with **Id = 4324234** (just as an example) is presented in the table below.

Table 6: Unlist aggregated devices, according to the number of aggregated devices

Id	Day	Hour	Speed	Number of Devices	Bin Geometry
4324234	26/01/2020	12	5,67	1	Point ((-9.123,38.755), (-9.1225,38,75615))
4324234	26/01/2020	12	5,67	1	Point ((-9.123,38.755), (-9.1225,38,75615))
4324234	26/01/2020	12	5,67	1	Point ((-9.123,38.755), (-9.1225,38,75615))
...	...	...	...	...	...

Due to different geographical asymmetric characteristics in our AoI (see Section 3), and considering the existence of soft mobility solutions just for one of the studied parishes 1, a comparison analysis was necessary for algorithm validation. The algorithm is expected to perform similar results in BSS stations location. Aiming the individual study in each parish, a geoprocessing operation was performed— this procedure allowed to delimit the geographic distribution of all docking points.

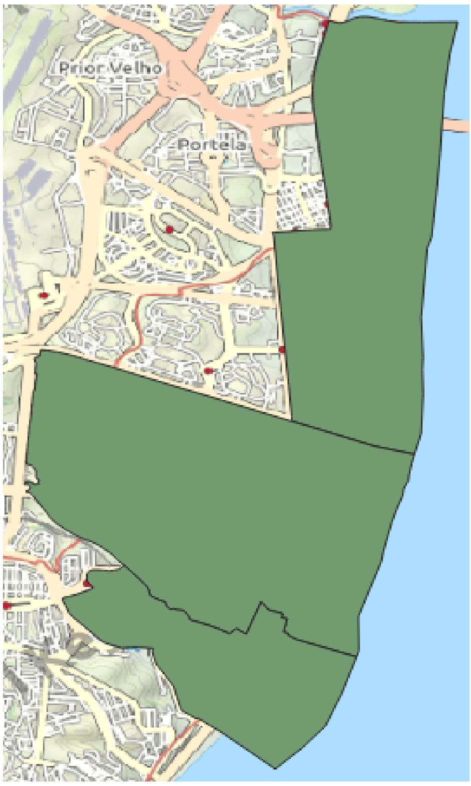


Figure 48: Area of interest of three council parishes

As presented in Figure 48, the area of interest needed to be separated according to the three council parishes, according to *Direção Geral do Território Portugal* (2020), so that it was possible to study each parish in an individual and suitable way, considering all the differences among them. For this operation, from a unique shapefile for the whole AoI, the individual shapefiles were then generation, resulting in three independent areas, as shown in Figure 49.

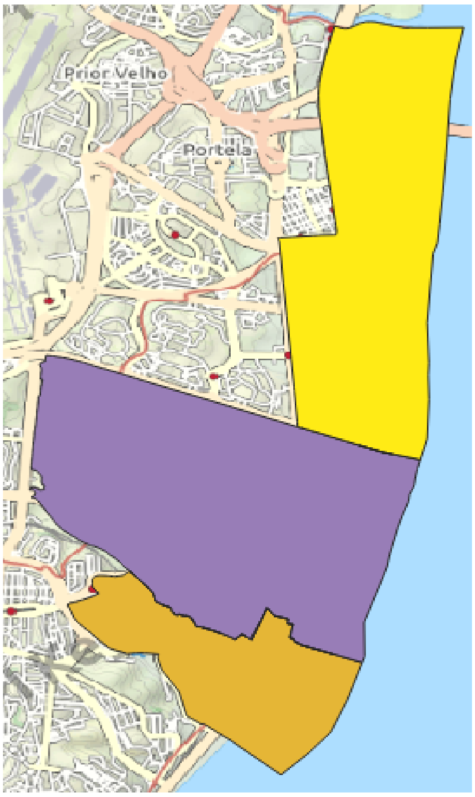


Figure 49: Individual generation of parishes limitations

After generating the individual parishes shapefiles of the AoI, then the **intersection** of **GSM data points** was performed, combining them with the **vectorial data** of **each** council **parish shapefile.**

After generating the individual parishes shapefiles of the AoI, then the **intersection** of GSM data points was performed, combining them with the vectorial data of each council parish shapefile.

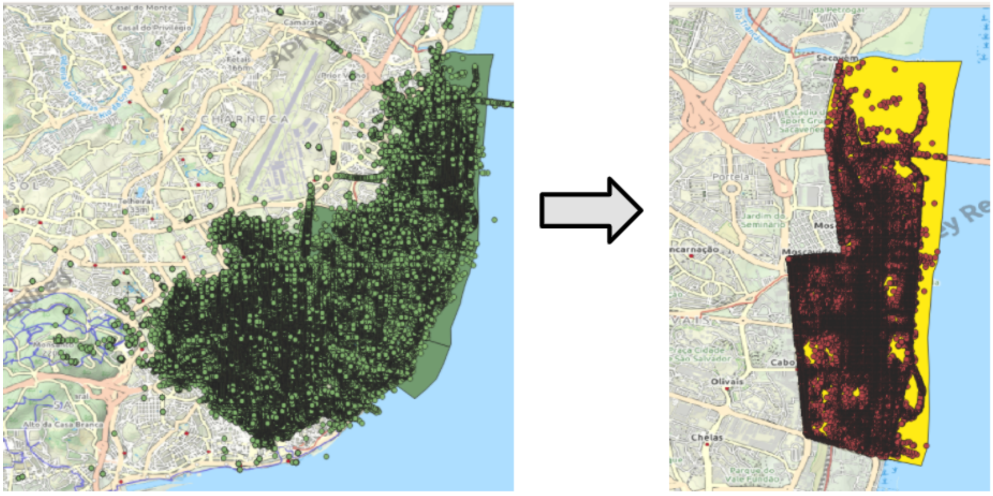


Figure 50: Intersection of GSM data points with vectorial data (shapefiles) - *Parque das Nações* case

For the development of the algorithm, an average velocity of 20km/h (see Section 5.2), independent of the user gender, physical condition, weather conditions, among others. Figure 51 shows a boxplot of the speed distribution after filtering it.

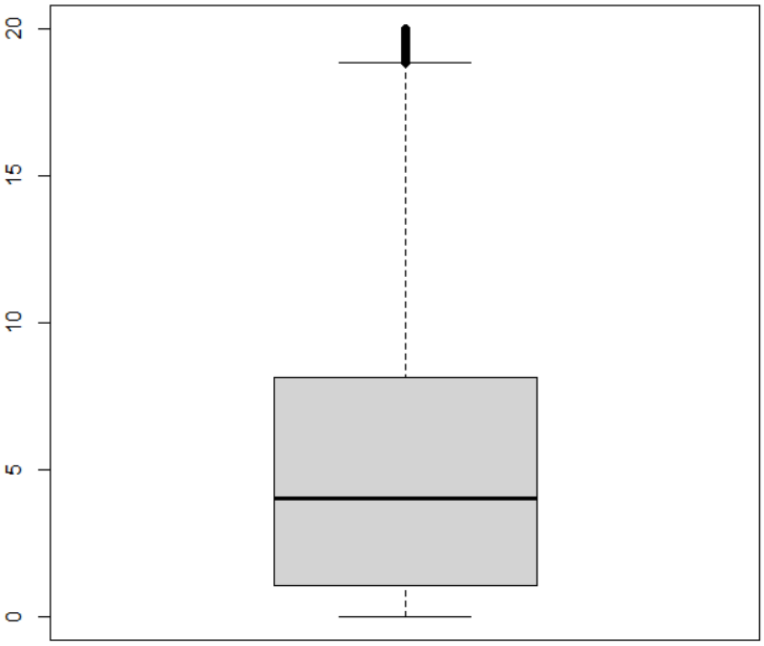


Figure 51: Application of filter -> collect samples with Speed <= 20km/h

This kind of filtering, only considering the samples with the column concerning **speed** under 20 kilometers per hour, is going to be possible to focus on several interesting devices distribution:

- Locate **Main traffic jams** - only considering low-speed devices will allow us to focus on the traffic jams that occur in our AoI, enhancing the definition of the clusters in the next step.

- Identify **Residential areas** - considering the night time frames, it will be possible to also provide and consider the data points that are referenced to residential buildings, social neighbourhoods, etc. These places must be important, as we saw in Chapter 4.

- Identify **workplaces** - in the daytime, low-speed detection might mean the geographical location of workplaces. This is very important considering the number of commuters that go to home-job and job-home.

- Focus on **soft mobility** solutions and **bicycle lanes** - filtering devices under 20 km/h allows us to consider also pedestrians and cyclists. This will be important to also collect important data points nearby or over bicycle lanes and close to existing docking stations in our AoI.

For this work, the **K-means algorithm was used.** This algorithm was chosen since the **main geospatial** clustering problems, as seen in Chapter 2, are typically solved with **partitioning** or **density** algorithms.

Moreover, and as studied in the mentioned Chapter, among the partitioning algorithms there are **important key** factors that influence the decision when it comes to select an algorithm to be applied. Considering the mentioned comparison between the K-Means, K-Medoids and CLARANS algorithm, available in Section 2.1.2, it was clear that **K-Means** was the least complex algorithm and also, one that offers **efficiency** when applied to larger dataset. Besides, it reveals a less complex algorithm when compared to the other two. For that reason, and considering the **mobility context** this dissertation is developed, the sensitivity to outliers (that can be harmful in most of the contexts), in our perspective, is important to analyse the outputs since **every data point** represent a **potential** bike-sharing system user, and for that reason, we should **suggest suitable location** for the docking stations that can **maximise** the operating radius to the most users as possible.

When compared to the density-based algorithms, if in one hand we have to define the initial value of K for the partitioning algorithms, in the density-based ones we have the two mentioned parameters, ϵ and MinPts. From our perspective, and considering the implementations found in the mobility domain, it was preferable to use well-known techniques to identify the K value for partitioning clustering over the **definition** of the **two required parameters**, in case of DBSCAN for example.

In this sense, considering that it was pretended to **evaluate** and **compare** the current distribution of the **existing docking stations** available in *Parque das Nações*, it was convenient the use of a algorithm which takes advantage of the **initial value** of **clusters**.

Hence, the first step when running K-means consisted of defining the number of clusters that should match the same number already installed of BSS docking stations. So far, there are fourteen docking stations in *Parque das Nações*. The same number of docking stations was considered when running K-means. In the development of this work, the Elbow Method was used for the identification of the initial value to be performed in the K-means algorithm for *Beato* and *Marvila*.

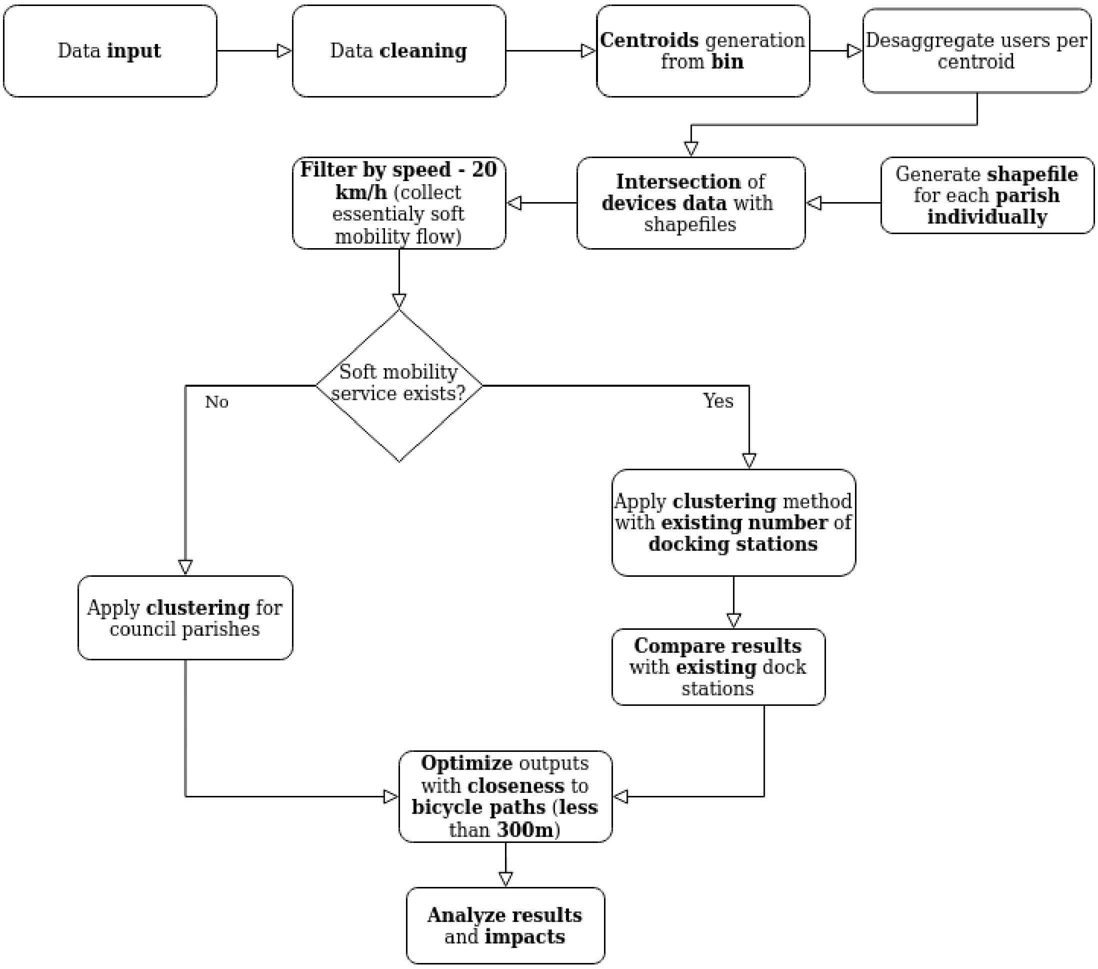

Data input
Data cleaning
Centroids generation from bin
Desaggregate users per centroid
Generate shapefile for each parish individually
Intersection of devices data with shapefiles
Filter by speed - 20 km/h (collect essentialy soft mobility flow)
Soft mobility service exists?
No
Yes
Apply clustering method with existing number of docking stations
Apply clustering for council parishes
Compare results with existing dock stations
Optimize outputs with closeness to bicycle paths (less than 300m)
Analyze results and impacts

Figure 52: Methodology applied for dissertation development

In order to evaluate results, especially in *Parque das Nações*, it was necessary to compute the distance between the in-site docking stations and the clusters output from the clustering algorithm. Hence, aiming at the calculation of distance between the clusters and the in-site docking stations, the **Haversine Formula** was used.

The Haversine formula, as described by Hartanto et al. (2017), is perhaps the first equation to consider when understanding how to calculate distances on a sphere. The word "Haversine" comes from the function: haversine $(\theta) = \sin^2(\theta/2)$ The following equation where ϕ is latitude, λ is longitude, R is earth's radius (mean radius $= 6,371$ km) is how the above formula is translated to include latitude and longitude coordinates. Note that angles need to be in radians to pass to trig functions:

$$a = \sin^2(\phi B - \phi A/2) + \cos\phi A^* \cos\phi B^* \sin^2(\lambda B - \lambda A/2)$$
$$c = 2 * \operatorname{atan} 2\left(\sqrt{a}_t \sqrt{(1-a)}\right)$$
$$d = R \cdot c$$

Thus, the Haversine formula has been applied to the pair (C1, DS1), where **C1** represents the **cluster output from the algorithm**, and **DS1** represents the **nearest in-site docking station installed**.

Noticing the importance of this step, the Haversine formula was also used for the optimising process of docking stations, as described in Section 5.3, following a consistent Workplan and being possible to correctly interpret results.

5.2 CONSTRAINTS AND ASSUMPTIONS

As any research project involves, some **constraints** are faced during the process and, even several efforts are put into action to mitigate their effects, they must be assumed and discussed in order to better understand the followed methodology and, consequently, the obtained results.

Thus, considering a real problem investigation, alongside the use of real data provided by telecommunication operators, there are some constraints, especially regarding data which are presented below. After a brief explanation of the several constraints, a short explanation for them is going to be provided.

- **Temporal Distribution** - considering the real data dump provided and the thousands of hundreds of records provided in the dataset, the temporal distribution of data was a clear limitation, which was known since the data was released by the operator. For that reason, and also due to legal obligations, only a small-time interval was possible to work with, in this case, month **January 2020**;
- **Temporal Resolution** - **limited temporal resolution** since all records were provided within an **hourly aggregation**;
- **Spatial Aggregation** - due to the GDPR reasons and imposed limitations in Portugal but also in Europe, the operation only could provide aggregated data for this thesis purpose, which introduces an important geospatial constraint since this project handles geographical planning for mobility and takes advantage of GIS data;
- **Geospatial aggregation methodology** - in order to take the previous point respected, the aggregation process followed a common Geometry Library denominated S2 cells, geographic markers used to map the Earth's surface, as explained in Section 5.1;
- **Multiple Device Consideration** - since the provided data is completely anonymised, each record has a unique identifier (Id), but that is not possible to have for each device (due to privacy and legal constraints). For that reason, and considering the geographical and temporal aggregations described here, it is possible

(and highly probable) that the same devices are going to be considered in different time frames and locations, and for that reason, **calculation** or **evaluation** of **number of devices** is **not** a **precise** metric to be considered

- **Mean Speed** - since the devices' information was aggregated in several records, this means a **loss of information** because **several data points** are now considered as **one**. For that reason, in specific metrics such as speed are affected and, for that reason, the **speed** considered is representative of the mean speed registered for all the devices considered in the aggregation, which affects the precision of any further consideration;

Hence, keeping in mind that several temporal and geographical constraints were faced in the project resolution process, it is expectable that the results obtained, described in Chapter 6 must consider that a small portion of data was considered in terms of time distribution and for that reason the results cannot be interpreted as a year of data was used(for instance, devices distribution and mobility patterns are affected for external conditions as weather, so that January vs summer months must present major differences).

Additionally, the kick-off for spatial representation of devices was S2 Cell, which means a polygon of approximately 10 meters square. For that reason, the **loss** produced in the **aggregation process** (converting raw data to S2 cells, and here the associated error to GPS must be also considered) and afterwards converting each **S2 cell** for its **centroid** must be bear in mind since one of the evaluation metrics of results is the distance between clusters output and in-site docking stations for *Parque das Nações*.

Furthermore, alongside several constraints faced, some important assumptions were taken into practice since they represented a way of minimising constraints effects and consequences. Some important considerations and assumptions are listed below.

- **Centroid Calculation** - in order to reduce the geographical complexity of the problem, since we had several polygons defining a group of aggregated devices, a post-processing method was needed in order to convert the geographical information in a simpler form keeping its meaningfulness. For that reason, the centre (middle point) of its cell was calculated, as it was assumed as a reliable way of producing the pretended effect.

- **Soft Mobility solutions and Speed** - as described in Chapter 2, soft mobility solutions are those that can be used just with human beings energy production. For that reason, and considering the objectives for bike-sharing docking stations identification, in order to focus this thesis on soft mobility solutions, we have assumed that the **speed** of **20 km/h** was a reasonable filter to **collect** mainly **people walking** or **stopped** (schools/business buildings, etc.). However, at the same time, due to the constraints mentioned previously, this filter will also capture **traffic jams** and mainly **congested roadways** in the city, which is an important fact to point out since it will allow us to identify clusters in some critical accesses in Lisbon.

- **Suitable distance from Bicycle Lanes and Docking Stations** - as studied in Chapter 2, BSS are implemented in several communities since 1965,when Amsterdam implemented what is now considered the first bike share system in history known as *Witte Fietsen* ("White Bikes" in English). Since that time, many BSS have been implemented across the world and the **distance** of **300 meters** have been used among them for bike-sharing docking stations planning, as it is seen as **reasonable distance** someone is willing to walk.

5.3 OPTIMISATION PROCESS

After the clusters generation process, the final step of the process mentioned above, an optimisation process was taken into consideration in order to better define the location of the new bike-sharing docking stations identification. This is very important in order to maximise the use of these systems, as seen in Chapter 2.

The optimisation process that is described in Algorithm 1 was implemented considering the state of the art guidelines for BSS, which in includes the proximity between bike-sharing docking stations and bicycle paths.

This process considered the direct output from K-Means algorithm and the bicycle lanes map from Lisbon city centre that belongs to each parish in the current study.

Hence, it was important to provide a particular distance for that optimisation, in this case **300 meters**. Usually, this distance is taken as a kick off for BSS implementation and other important geographical planning problems in mobility context.

Algorithm 1: Optimization for bike-sharing docking stations

```
Result: List of docking stations nearby bike paths
Apply K-Means algorithm on specific council parish;
Define threshold for otimization (in meters);
For each centroid(c) output from k-Means do
    For each segment(seg) of bike-path do
        Project c in seg, output is np, following nearest point definition;
        Calculate distance(d) from np and c;
    If the minimum distance from d to np is lower than threshold;
        Adopt the point in segment as optimal point (located in bike-path);
    EndDo;
EndDo;
Print optimised points;
```

6

DISCUSSION AND RESULTS

In this section, we are going to present the results that have been reached, following the methodology presented in Chapter 5. First, the results regarding *Parque das Nações* are going to be presented and analysed, comparing the algorithm output with the existing GIRA bicycles docking stations already implemented. After, an extensive analysis to obtained results for *Marvila and Beato* is presented, **highlighting** the **number** of d**docking stations suggested** and, most important, **suitable location** for those docking stations

6.1 PARQUE DAS NAÇÕES

As presented in Chapter 3, the current bike-sharing system implemented in Lisbon city has fourteen docking stations available in *Parque das Nações*. For that reason, the K-Means Clustering Algorithm was applied to the prepared and clean data after the process mentioned in Chapter 5, with the 14 clusters as output expected.

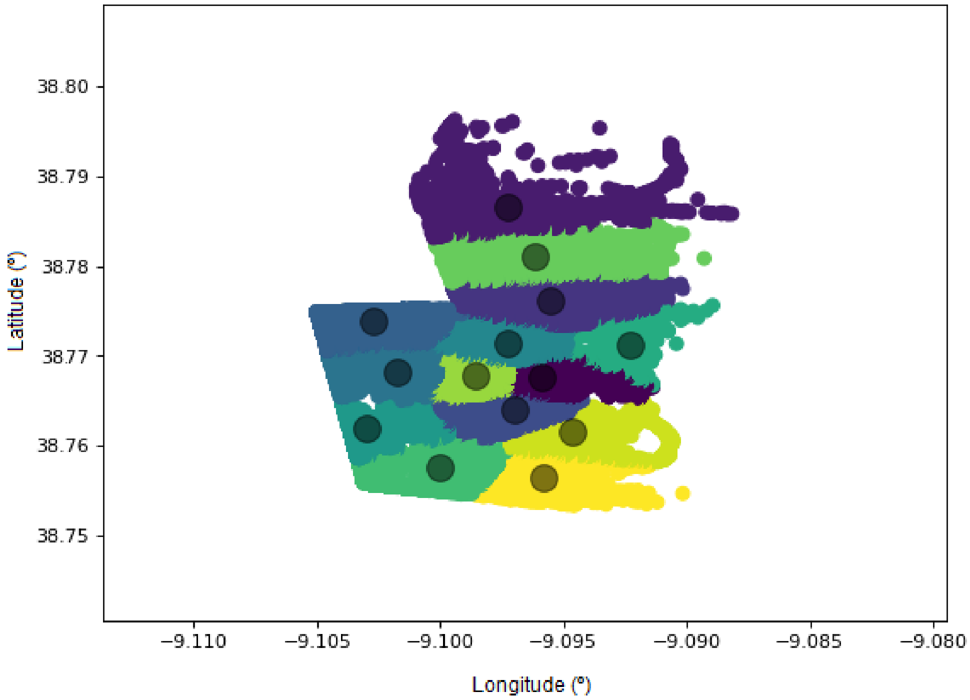


Figure 53: K-Means centroids output for *Parque das Nações*

As Figure 53 suggests, there are particular areas where the centroid generation reflects a harder process of clustering since we can see a higher dispersion of devices, namely in cluster number 1 (from top to down perspective.)

However, this representation is harder to identify the reallocation of the clusters, with the axis x and y representing the variation in longitude and latitude, respectively. Figure 54 illustrates the K-Means clusters distribution, considering their area of application.

The map was divided into three sections showing the differences between **existing docking stations** (Gira Lisbon Bicycles) and the followed methodology. Analysing the data in detail, it is possible to verify that there are fourteen points for both cases.

From a comprehensive perspective, most yellow points (existing docking stations) are next to dark blue points (optimised clusters). In those dots, the algorithm has worked as planned (having a difference under 200 meters, mostly). Therefore, only the dark blue dots away two hundred meters for the yellow dots will be analysed in detail. Note that the red dots represent the **direct output from the K-Means Clustering**, with no optimisation.

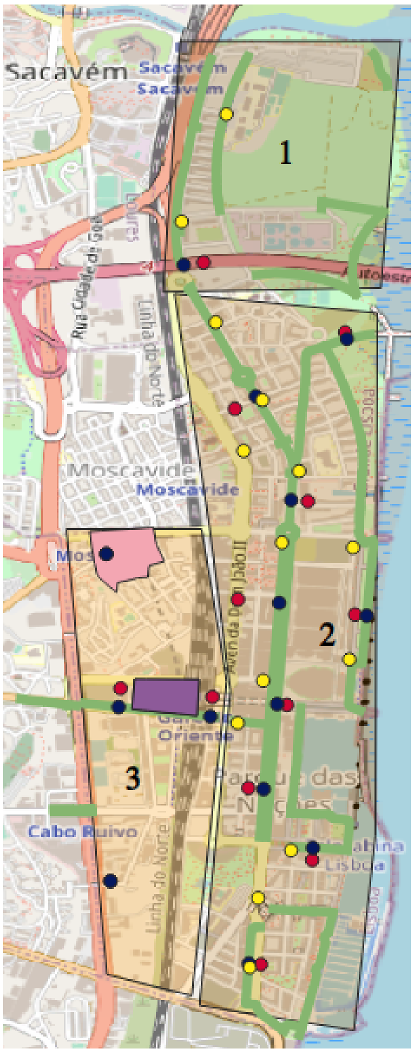


Figure 54: Bike-sharing docking stations in *Parque das Nações*: Existing docks in-site are represented by yellow dots; red dots represent centroid clusters; dark blue dots represent the optimised locations for k-means outputs (docking stations over bike paths). Subway has the pink colour polygon while the purple polygon regards the train station. The map was divided in three segments.

In the **first area**, **GIRA bicycles** have implemented two different docking stations over two segments of bicycle lanes in this area. However, we have identified this area as an interesting section for analysis and discussion due to these**particular attributes**. In this area, there are a considerable "inaccessible area" due to existing buildings: a huge fenced terrain which takes a huge part of the mentioned area, and also, a Waste-Water Treatment Plant (WWTP), which obviously, is not accessible for common mobile phone users. For that reason, this area has shown particular features which have turned into a specific sub-case inside the *Parque das Nações* analysis area. Figure 56 proves the existence of such infrastructures.

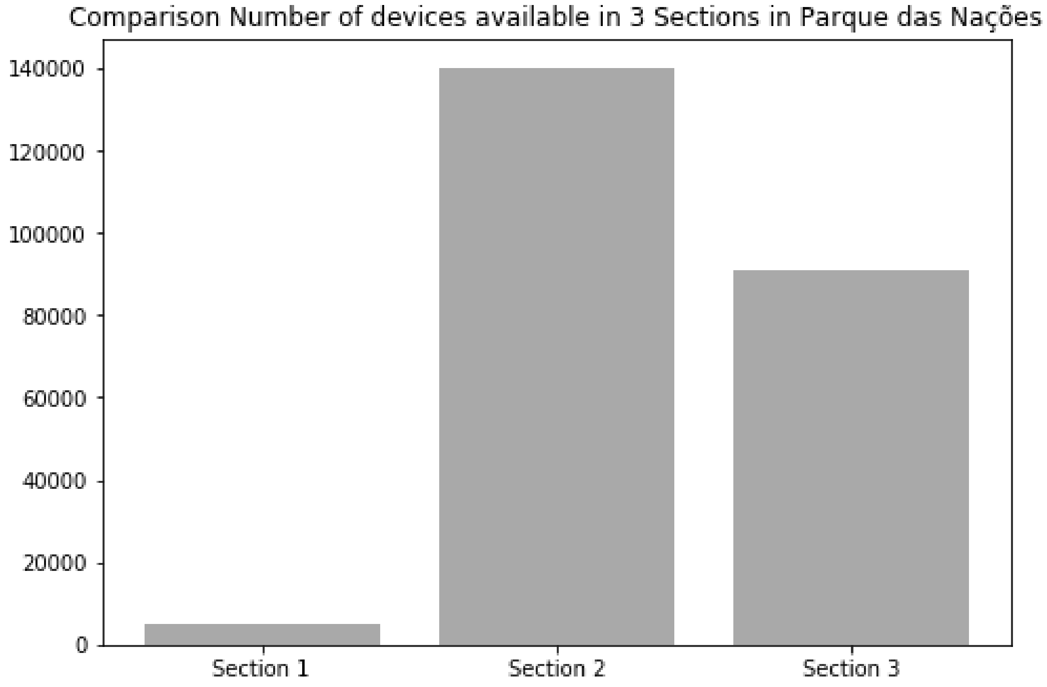


Figure 55: Comparison: Number of Devices in the 3 different areas/sections in *Parque das Nações*

For this reason, our **our cluster algorithm** has **our cluster algorithm one docking station as needed**, considering the number of devices and its spatial distribution for the whole parish (keep in mind that the clustering method was applied to the whole council parish area). The identified docking station (the only dark blue dot in the first section) is close to several green spaces (*Parque Tejo* or *Jardim do Arco da Expo*, for example) or some leisure places such as *Skatepark Expo*. Nevertheless, the **data** used for clustering belongs to **January 2020** as mentioned before, a winter month. For that reason, in addition to the features of the first section, the **concentration** of the **devices** in that area might be affected by the **weather conditions**, creating a **natural bias** of **lower devices** distribution in that area.

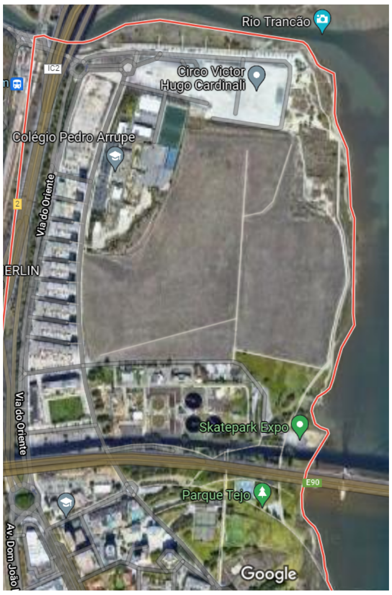


Figure 56: Satellite imagery from Google Maps for Section 1 (*Parque das Nações*) in 2021

If remind the analysis conclusions in Chapter 4 for *Parque das Nações*, we can assume that the algorithms' output proofed our initial predictions. For *Parque das Nações*, the distribution of the devices alongside the first section was very residual. For that reason, considering that we have expected the **fourteen most concentrated** in that region, we find it acceptable to identify only one docking station in that area.

Regarding the **second area**, we have the major number of docking stations identified. This kind of event can be justified by some interesting geographical reasons, such as:

- **Closeness** to the **river** - this part of the map is alongside the west bank of the Tagus river. Due to this, several maritime activities are then located in this area, such as *Porto de Lisboa*, for example.
- **Commercial buildings** - in area number two, as marked in Figure 54, we can find some attraction PoI such as restaurants and *Vasco da Gama Shopping*. Due to the existence of such attraction services, the amount of data points there is much higher than first area (in the north), for example.
- **Leisure Places** - as reference above, it is intrinsic the closeness of area/section two to the river. Consequently, several activities which often depends on water can be found there, such as *Lisbon Oceanarium* and Lisbon marina. Additionally, we can find some EXPO 1998 buildings too.
- **Workplaces** - as studied in Chapter 3, in this area we can find several offices and, for example, In this sense, these surrounding areas are more often affluence during the day, causing the common known rush-hours, but during the night the devices concentration gets much lower.

Described several reasons for the higher number of docking stations identified in area 2 of Figure 54, it is now time to visually analyse our outputs. In a general observation, we can verify that **dark blue dots**, i.e. our **findings**, are very **close** to the **existing docking stations** (**yellow dots**). That means that our **workplan** and **followed methodology represents** a **reliable** and **accurate** approach to find **soft mobility hotsposts**, in particular, BSS docking stations. However, it is still possible to identify an **isolated dark blue** point, which has no close neighbors. After noticing that fact, we have took a look for the existing nearby services and other PoI that might justify that identification, apart from one of major factors in mobility: **bicycle lanes closeness**. This **upper east** point is the nearby a juvenile garden and very close to Tagus river, and also it is also linked to the existing bicycle lanes available in *Parque das Nações*. For that reason, this **hotspot** is seen as a **serious** and **potential candidate** for further additions in the current system operating in Lisbon.

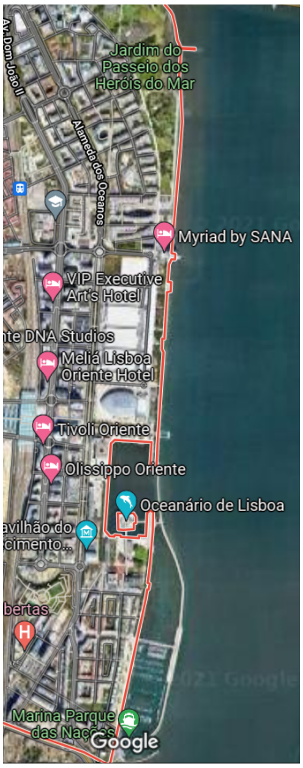


Figure 57: Satellite imagery from Google Maps for Section 2 (*Parque das Nações*) in 2021

Lastly, we have now **area/section three.** This area, as described in the map, is **very important** because it contains several interesting conclusions that are possible to be drawn. First of them, is the fact that **this regions** is **currently uncovered** by the GIRA bicycles system. In addition, **four** out of **fourteen** clusters are located inside this area, meaning that **30%** of our findings for *Parque das Nações* **highlights** the **need** to **increase** the **geographical coverage** of the current system in this council parish.

Taking a look to the exact location of such clusters (dark blue dots in area three), we can express the **correlation** between the train station in *Moscavide* (ponk polygon) and also the train and simultaneously subway station *Gare do Oriente*, which aggregates **three out** of the **four** dots in this area. This fact **underlines** the importance of locating docking stations nearby these kind of mobility PoI, regarding such agglomerations of devices found. For

Gare do Oriente, in specific, we can see two docking stations identified for each side of the station, pointing out the **massive number of devices** nearby this location, as we have already analysed in Chapter 4.

To study the identification of the lower left side dot of the section (the Southeast one), it was unveiled the location of such point: *Avenida Infante Dom Henrique*. As explored in the Data Exploration Chapter (here 4), this is one of the most important roadways in Lisbon city center and moreover, in our AoI. Since this avenue crosses the all three parishes of study, and as seen previously the such traffic jams and congestions within this access in *Parque das Nações*, the identification of this "lonely" dot provides a **key role** for future improvements in this system. As an increase, the location of the latest point is in accordance with the public plan released by *Câmara Municipal de Lisboa* to extend the current bicycle lanes network, as shown by Motor24 (2019).

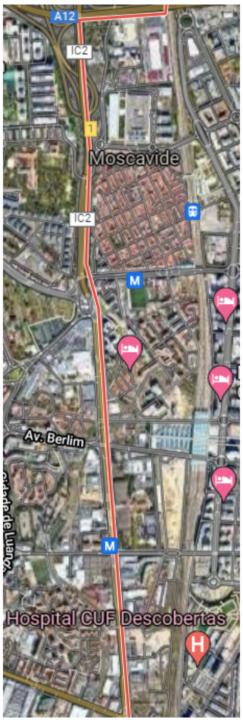


Figure 58: Satellite imagery from Google Maps for Area 3 (*Parque das Nações*) in 2021

In order to measure how far the centroids output distance from the existing bike-sharing docking stations, an evaluation algorithm was applied. For each centroid, the closest GIRA docking stations was then identified, and after that, the **Haversine distance** (explained in Chapter 5) between the two coordinates is computed. The global results are presented in Figure 59.

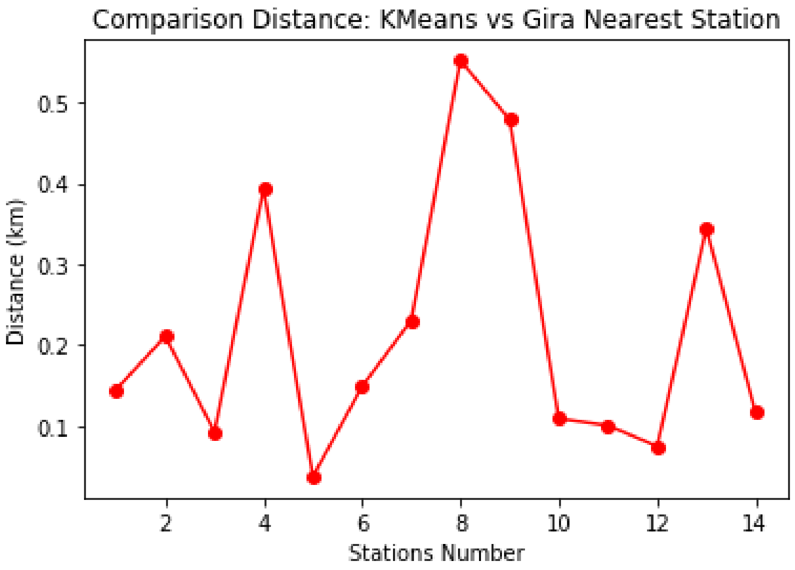


Figure 59: Distancing comparison: K-Means output vs Gira bicycles docking stations

As we can see in Figure 59, our **results** are, in its majority, **very close** to the existing docking stations. As we can see in the plot above, we have identified devices concentrations nearby the existing docking-stations in *Parque das Nações*, corroborating which was shown in Figure 54.

From Figure 59 it also possible to extract other important conclusions regarding this evaluation metric. One important insight is the fact that there are many docking stations identified very close to the existing ones. For example, **six** of the **predict docking stations** are distancing under **200 meters** from the existing ones, as shown in Figure 60 being here a perfect match between the algorithm output and the existing docking stations. In this sense, we can concluded that **area/section two** produced the best results, according to this plot.

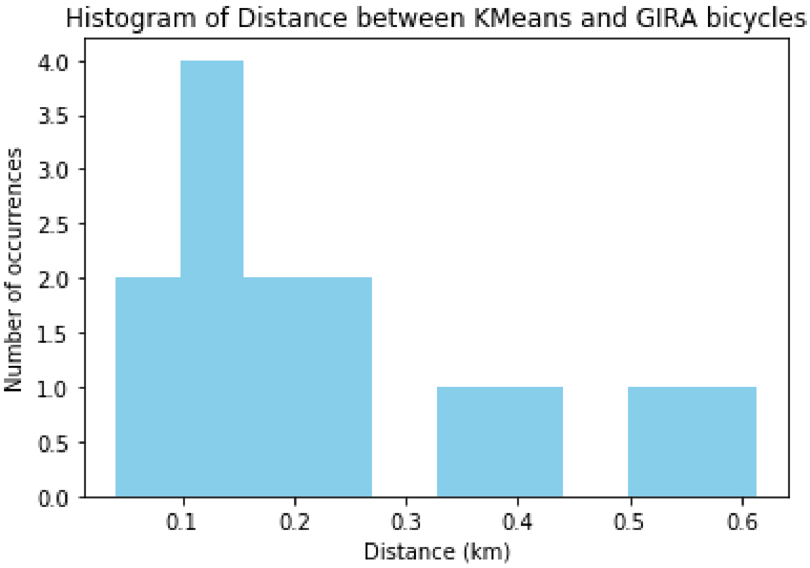


Figure 60: Histogram of Distances comparison for KMeans and GIRA stations

On the other hand, the most away stations in *Parque das Nações* output belongs to area three, being the northeast and the Southeast points of this area the most far away from the existing ones. However, this kind of **results** does not mean any kind of **poor performance** from the algorithm and the methodology, since this has been studied the reasons and, as mentioned, they represent the **identification** of **new docking stations** in **uncovered area**.

An important note for this evaluation process is the fact that the Haversine distance have considered every clusters output from the algorithm and **compared** them with the **nearest docking station** from the current system implemented. Due to this fact, this analysis might consider more than once the same docking station. After verifying this, we have concluded that **the repetition** of the **nearest docking stations** occurred mainly for the section three, where the system is not implemented yet.

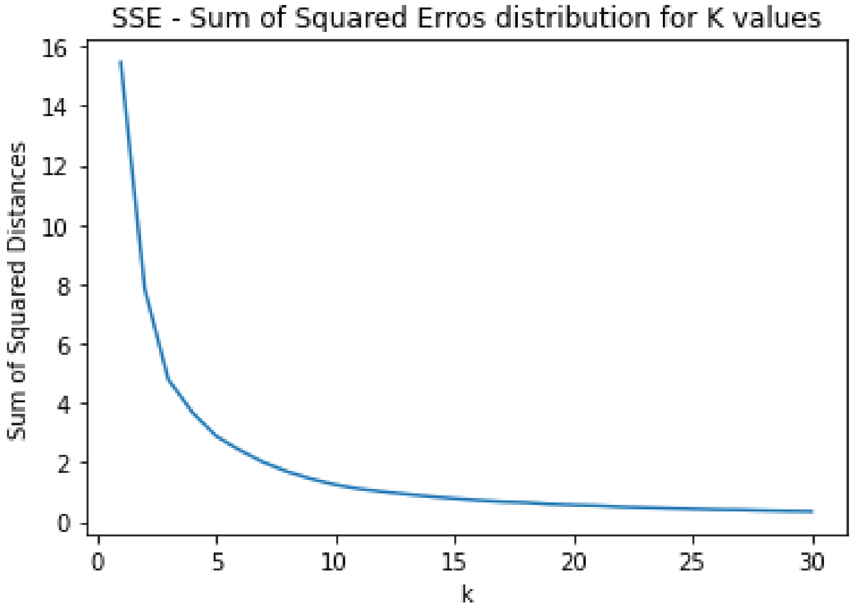


Figure 61: SSE Plot for *Parque das Nações*

From Figure 61 it is possible to verify the variation of the Sum of Squared Errors according to the number of clusters (k) defined. As stated in Chapter 5, and also as described and shown in Chapter 3, Lisbon has a public bicycle sharing system operating in *Parque das Nações* with fourteen docking stations on-site at the moment.

As the plot suggests, the **current number** of **docking stations available** does not match with the Elbow Method that is used for *Beato* and *Marvila* cluster's number definition. That said, it is extremely important to consider the mobility context we are studying on, and for that reason, enhance the need oh having a **wider** geographical coverage in order to maximise the efficiency and usage of such system, as reported in Chapter 2.

From the direct observation of the SSE plot, we could suggest the "optimal" number of K for this council parish is between **5** and **10**, being extremely close to K = 7 the elbow from the SSE curve. Nevertheless, it is known that there are fourteen operating docking stations, which indicates that the Elbow Method **must be** an **indicator** of **the minimum number** of **docking stations** for the amount of data provided, and considering the dispersion such clusters are going to suffer.

Considering the 14 docking stations, it is evident that K = 14 means a further stage of SSE distribution, getting more flat with this value. This is also an indicator for the upcoming analysis of this metric, since *Parque das Nações* is our **baseline** of **work** for the other two council parishes.

From a real world interpretation, the **SSE analysis** reflects the **"difficulty"** of **aggregating** the initial data points into the given number of clusters. Hence, this expresses "how far" are the final clusters output from the given data points that were used to generate that cluster. Combining that information with the real distancing between docking stations and clusters output in Figure 59, K = 14 suggests, in terms of average values, **distances under** 600 meters from the real locations, being the majority of the gaps lower than 400 meters.

6.2 BEATO AND MARVILA

Beato is the smallest council parish in analysis in this study. Nevertheless, it still provides some interesting/crucial urban mobility flows, and it is possible to identify exciting spots for bicycle docking stations.

In order to identify the location of new docking stations in *Beato*, an additional step was taken into account: **definition** of **number** of **docking stations** that should be installed in this council parish.

For that reason, the K-Means algorithm has been applied for several K values in order to determine the optimal K, i.e., the most suitable K value for *Beato* data distribution. In Figure 62, the distribution of Sum of Squared Errors (SSE) for the respective K value is presented, in *Beato*'s case.

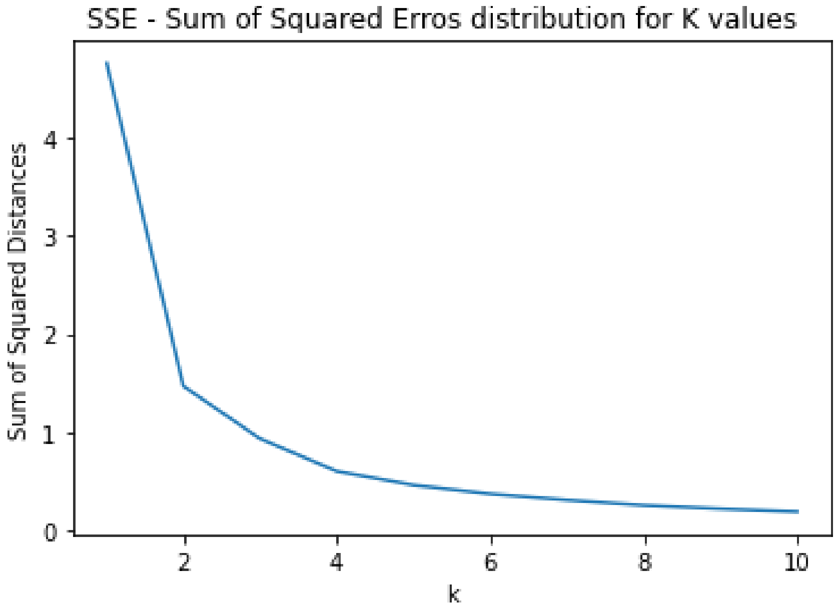


Figure 62: SSE variation according to K values in *Beato*

As seen in Chapter 2, one of the most well-known methodologies for the identification of the best k value in Clustering problems is with Elbow Method. So, analysing the SSE plot for K values in Beato, we can assume that the k = 3 represents the "Elbow point", since at this value the **error decreases** in **significant way** , in comparison with the error value. However, as discussed in Figure 61, the stabilisation of the error value is something that is important, considering the spatial distribution of devices and also the context of this problem.

In this sense, we can see from the plot above that the error gets **lower distortion** when K = 4. At this point, the SSE value gets less than 1, which indicates now a **very significance** reduction in this metric value if consider K = 2 or K = 3. Hence, since the value is getting lower, that means **cluster identification process** is getting **more**

"accurate" considering dataset. That means, since it can be used as a **measure** of **variation within** a **cluster**, and considering the K-Means algorithm **assigns** all data points to the **closest center** based on their **Euclidean distance.** Since considering a geospatial problem, it is measuring the deviation from the **centroid** to the **data points** have been used for its generation. So that, and according to the **measurable results** obtained in Section 6.1 that means, if choosing a **minimum number** of **docking stations** to be installed as K = 5, this represents an SSE of 0.462912. From *Parque das Nações* experience, it was described the **average** distancing for the SSE value in this council parish, so based on a similar reasoning, here in *Beato* we are obtaining, in general way, half of the SSE when compared to *Parque das Nações*, which means **obtaining centroids** distancing, in average, **200/300 meters** from the real location of devices, which is **perfectly aligned** with some of the generic guidelines for BSSs, as seen in Chapter 2.

Regarding *Marvila*, the reasoning applied was very similar to the one in *Beato*. Figure 63 illustrates the variation of SSE for for this council parish.

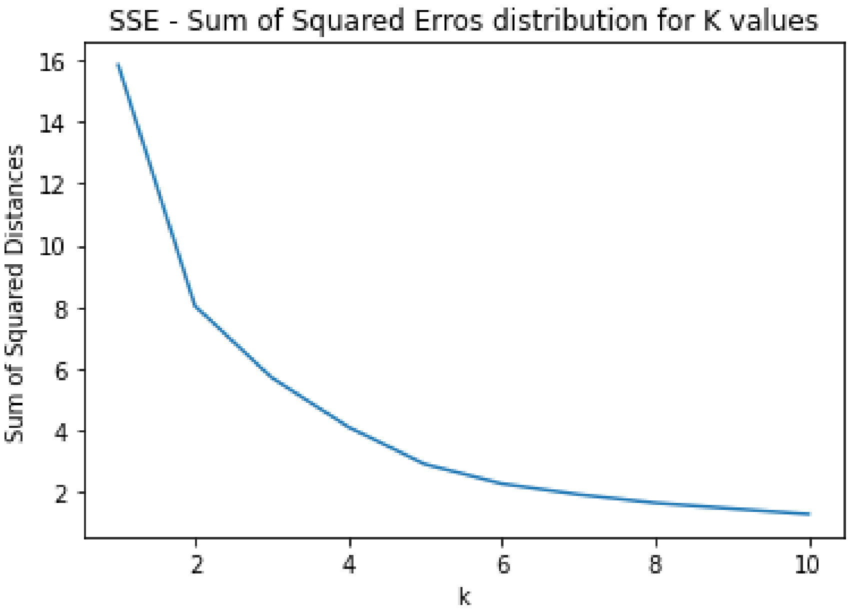


Figure 63: SSE variation according to K values in *Marvila*

Following the same argument as stated for *Beato*, as should be expect, the SSE reflects a curve with a **sharp decrease** as the K values increase. Moreover, it is still possible to suggest that the "Elbow point" for *Marvila* is somewhere between K = 3 or K = 4, but is the corresponding SSE value is **very high**, considering the physical representation of such values in terms of **distancing** from the **devices location.** In this sense, and taking into account **real world context** and **urban mobility planning** issue, it is suggested to consider a **higher K value** in order to that way **decrease** the value of **SSE** to a **similar scale** to the **previous ones**,i.e., for the other two parishes. Nevertheless, it is notorious the **higher dispersion** of data (better shown in 65), and for that reason, here it is output higher values from the SSE. The discussion for this behaviour is going to be discussed next.

Thus, in order to **keep available** a **sparse** solutions regarding the docking stations installation, and following the same concept of **SSE stabilisation**, it is suggested considering **K = 8** as a **minimum number** of docking stations to be installed in *Marvila*, and still outcomes an SSE error of 1.662275, which represents a **higher distribution** of of devices (data points), when compared to the **centroids generated**. This number (K = 8) have been chosen following the same criteria before, pointing for a flat stage of the metric in analysis, and also to **minimise** the **effects** of a **higher geographical distribution** of points.

From a visual perspective, Figure 64 illustrates the distribution of data points for *Beato* and the centroids obtained.

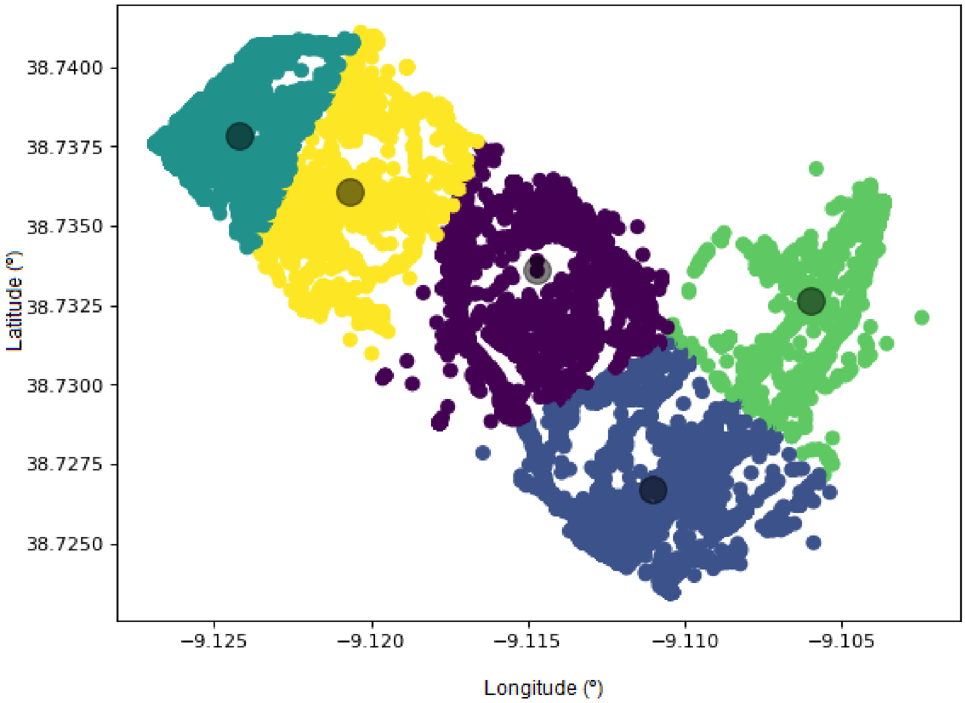


Figure 64: K-Means centroids output for *Beato*

As detailed before, for *Beato* it is suggested the addition of **five** new docking stations. If look at **X** and **Y axis**, which represents the **longitude** and **llatitude**, respectively. Considering the variation from such axis, we can verify a that, for example, longitude variation occurs in **two hundredths**. Mapping this variation into the real world, it is possible to note the geographical distribution of the points is according to the natural boundaries of *Beato*, as shown in Figure 67.

Regarding *Marvila*, this council parish have been revealed as more challenging scenario. Considering a direct comparison with *Beato*, it is possible to verify a **significant higher** variance in **X** and **Y axis**, since *Marvila* has more arean than *Beato*. For instance, Figure 64 expresses a variance of two hundredths on longitude scale, whereas Figure 65 presents a variance of **five decimals**. Transposing such facts to the real world context, that is verified due to the **extension** of this two parishes, where *Beato* presents an area of 2,46 km^2, while*Marvila* comes with 7,12 km^2, according to de Lisboa (2021).

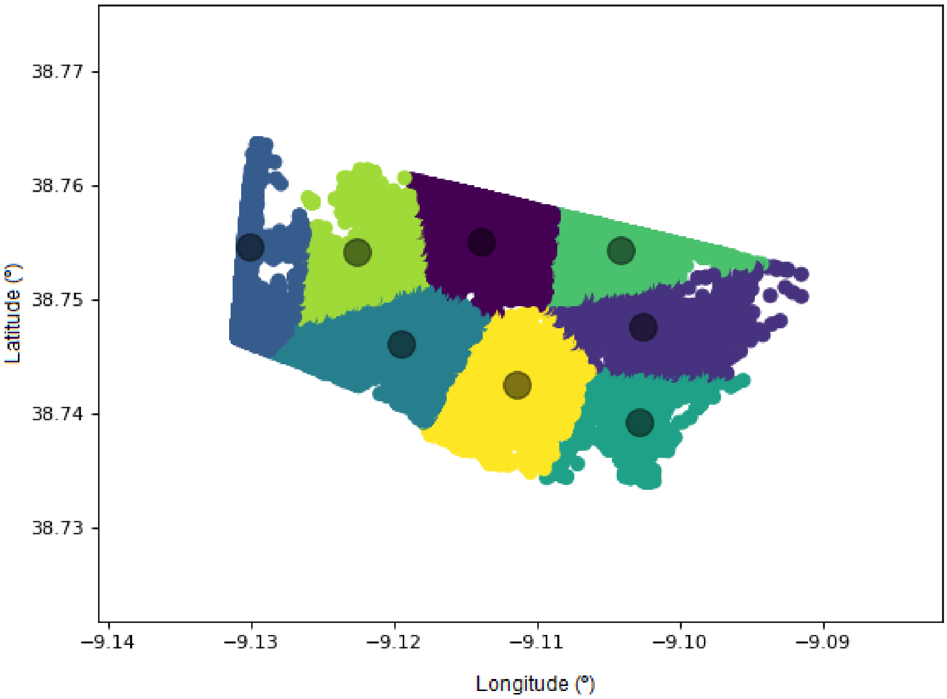


Figure 65: K-Means centroids output for *Marvila*

Considering the distribution of data points and clusters identification on Figure 65, it is possible to observe **data points distribution** is **confined** to the **geographical boundaries** of *Marvila*, as we can see in Figure 69, validating the previous process of **data preparation**. Alongside this conclusion, it is possible to call attention for the **balanced distribution** of centroids according to the points, being possible to admite they are in **central area** of **their neighbours**, which allows us to conclude they will provide **such positive** impact in the mobility context, since they are **equitably** distancing from **all the devices** nearby (and belonging to such centroid).

In a pragmatic point of view it is notorious that c**clustering** data points from *Marvila* is harder than clustering for any other parish from our AoI, which is reflected in the SSE plot and the **number of docking stations suggestion**, but also here, from the visual analysis of such clusters. Hence, even with a **balanced distribution** of the **clusters**, they are representing **wider dispersion** of geographical points, which can be justified by the **much higher area** from *Marvila*, when compared to *Parque das Nações* or *Beato*, and consequently, greater distance between the data points.

Corroborating the **chosen number of docking stations** as needed is also, the **amount** of **devices available** in each parish. Figure 66 illustrates the contrast between the three different parishes.

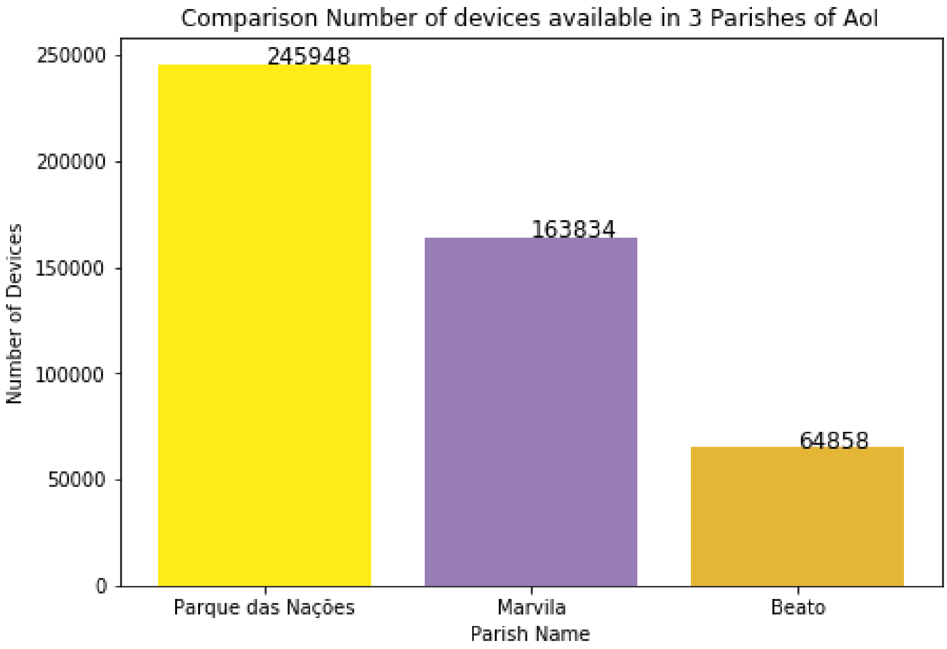


Figure 66: Number of devices considered for analysis in each parish of AoI

That said, it is important to underline that, even with a **higher** dispersion and **error** for *Marvila*, the **number of devices** within the dataset **does not suggest** the **need** of **increasing** the **number** of **docking stations** regarding this parish has **less devices** when compared to *Parque das Nações*, but they are **further away**, leading to the greater value of SSE, as explained before.

Considering this, and realising that the **definition** of **minimum number** of **docking stations** have taken into account the **geographical distribution** of **devices** and **number of devices detected**, this process implies a **tradeoff** towards the **geographical coverage** of provided datasets, but also the **potential usage rate** considering the installation of docking stations: *Parque das Nações* provides more devices when compared to *Marvila*, while *Beato* is the least agglomerated parish (according to GSM data provided).

After realising the **minimum amount** of **docking stations** set to install or add in the different council parishes, it is time to analyse the **suggested locations** for these soft mobility hotspots.

Concerning *Beato*, as documented before, there was found out that **five** docking stations would be a minimum number for this parish, according to the different criteria mentioned. The output is presented in Figure 67.

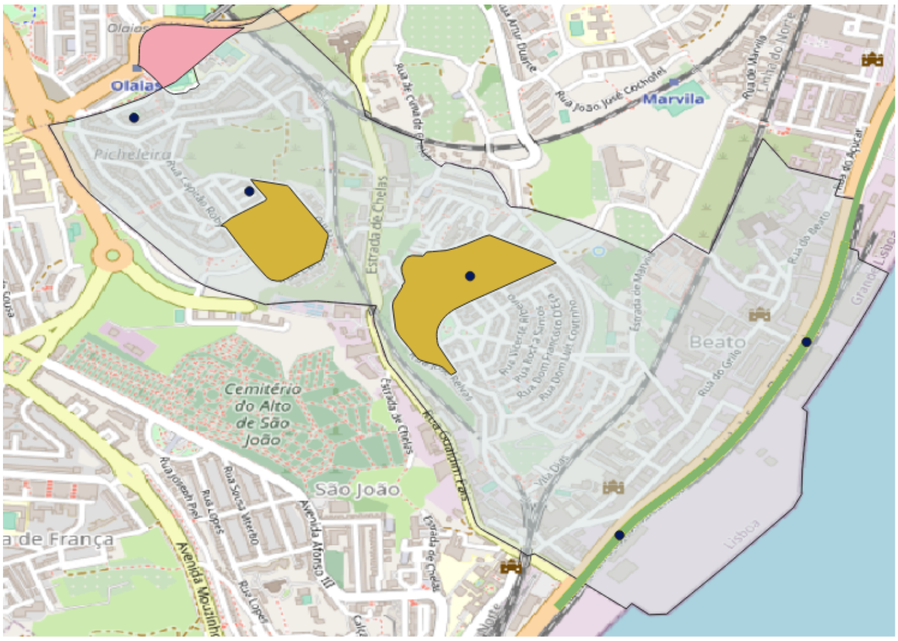


Figure 67: Bike-sharing docking stations prediction for *Beato*.Dark blue dots represent the optimised locations for k-means outputs (docking stations over bike paths). Subway has the pink colour polygon while the purple polygon regards the train station. Schools are coloured yellow and bike paths green.

As possible to see in Figure 67, our docking stations have already been **optimized**, since the two docking stations close to the Tagus river are already over the bicycle lanes available in this parish.

Here in *Beato*, Figure 67, the identification of hotspots have been corroborated by some insights we have concluded from *Parque das Nações*. For that reason, and since *Marvila* output shares mainly these characteristics, a detailed analysis is going to be made for both parishes.

For a better understating of parish geographical representation, an satellite image is provided in Figure 68.

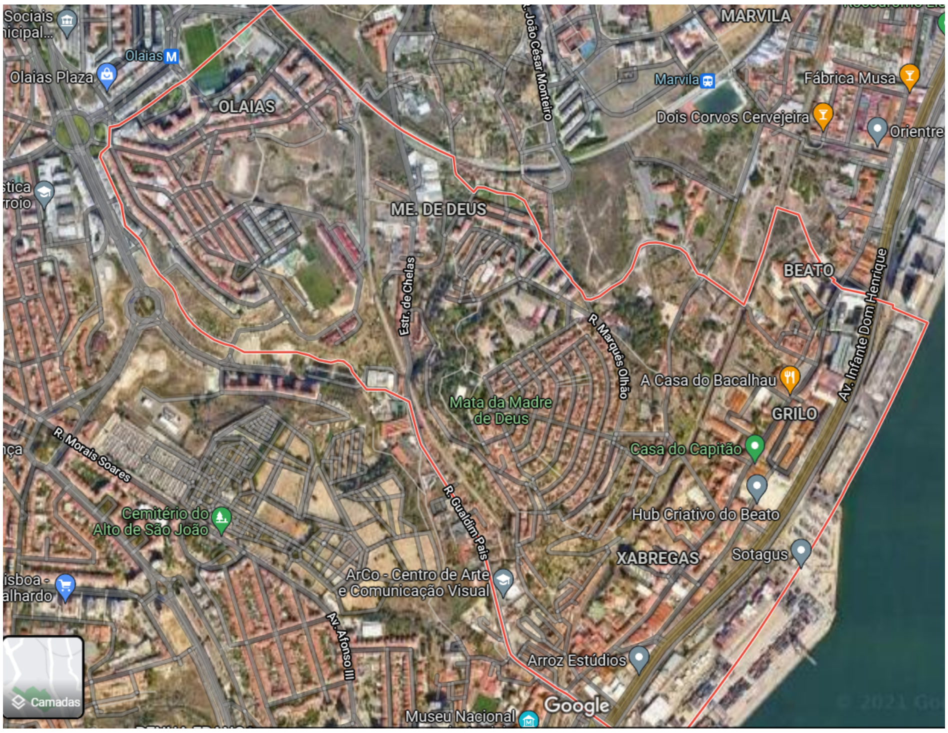


Figure 68: Satellite imagery from Google Maps for *Beato* in 2021

For *Marvila*, results are presented in Figure below.

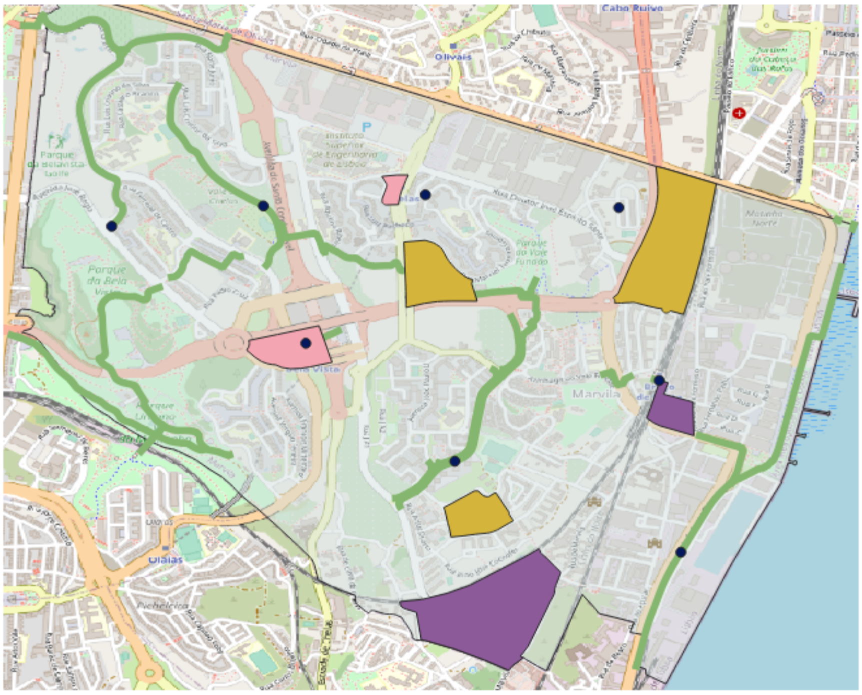


Figure 69: Bike-sharing docking stations prediction for *Marvila*.Dark blue dots represent the optimised locations for k-means outputs (docking stations over bike paths). Subway has the pink colour polygon while the purple polygon regards the train station. Schools are coloured yellow and bike paths green.

Firstly, from the observation of Figure 67 and 69, it is noticeable the importance of the mobility PoI in the agglomeration of devices. For instance, and considering *Olaias* subway station for *Beato* and *Chelas* and **Bela Vista** subway stations in *Marvila*, we can conclude the **high affluence** on such **areas**, also intensified in *Olaias* subway station due to having the *Olaias Plaza* shopping centre in the same location.

In same sense, educational PoI still represents influence on the clusters identification process, as there get populated by students but also from professors and staff for example, which represents usually hotspots of devices. In *Beato* and *Marvila* this is something possible to conclude, with *Escola Antonio Verney* or *Escola Básica Duarte Pacheco* for *Beato*, and for example *Centro Artes Marvila* and *Escola Básica de Marvila* being also identified.

In order to provide a more reliable view of this area, Figure 70 is presented.

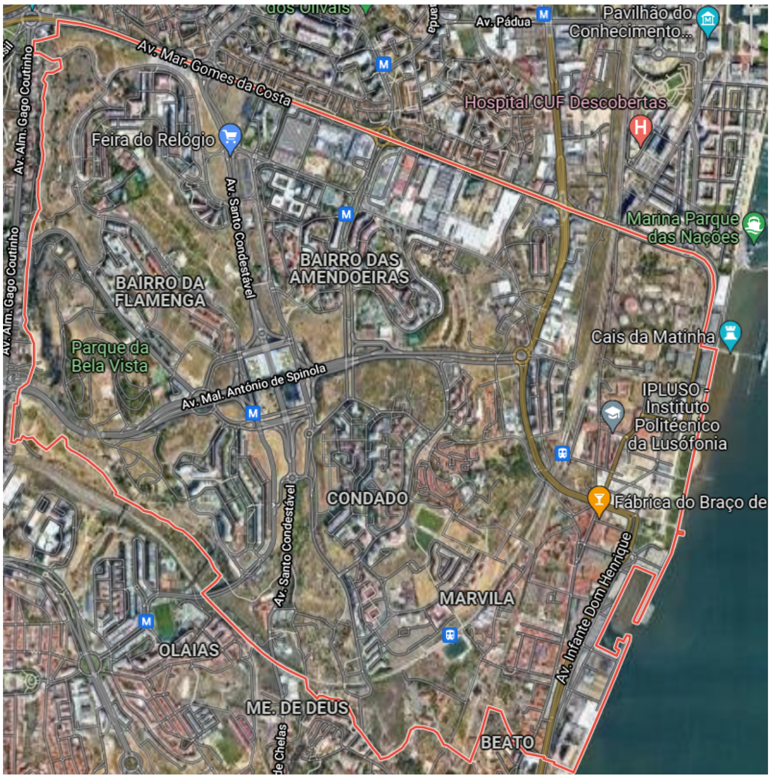


Figure 70: Satellite imagery from Google Maps for Section 1 (*Parque das Nações*) in 2021

One more point important to highlight, and also as mentioned in Section 6.1, in *Avenida Infante Dom Henrique* is usually face several traffic jams and frequently congestions, due to the **high number** of **cars stopped** in this avenue.

For a better understanding of these kind of situation, Figure 71 the devices distribution in this Avenue.

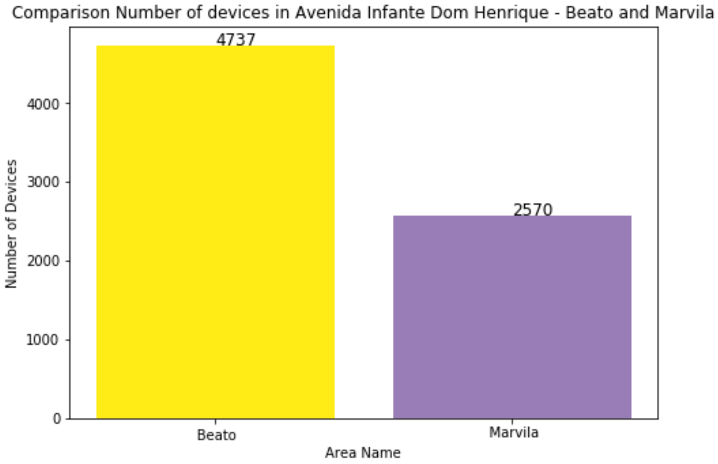


Figure 71: Devices Distribution in *Avenida Infante Dom Henrique*: *Beato* vs *Marvila* segment

As we can see from Figure above, the devices distribution for the segment of *Avenida Infante Dom Henrique* in *Beato* and *Marvila*, considering this avenues crosses both parishes. Thus, we can verify a **higher number of devices** in *Beato*, alongisde this avenue with 4737 devices detected, whereas in *Marvila* there was 2570 devices detected.

However, this number of devices are distributed in different affluences during a single day. It was chosen a **random weekday** for this analysis. The distribution of devices for the considered day is presented in Figure 72.

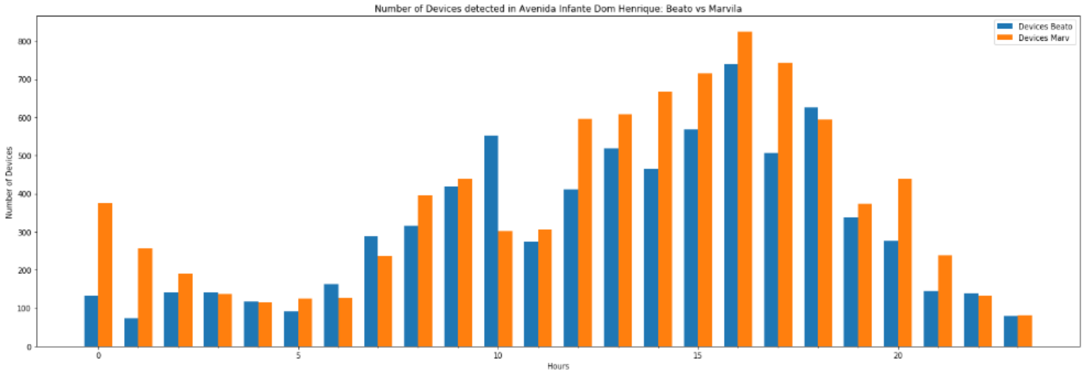


Figure 72: Devices Distribution in *Avenida Infante Dom Henrique*: *Beato* vs *Marvila* segment for each hour

As possible to see in Figure above, there are some **interesting hours** that reflect a higher concentration of devices. Namely **during the morning**, from **7am to 10am** it possible to **confirm** an **increase** of the number of devices circulating in that roadway, which considerable **higher number** of **devices** during the day, when compared to the night frame.

For that reasons, our clustering process have identified **three docking stations** as needed in *Avenida Infante Dom Henrique*, **allowing** to have **multiple docking stations** nearby the Tagus river, and also parallel to this avenue, being possible to pick or leave a bicycle in several points of this avenue. Additionally, this is a very important avenue since it has multiple traffic lanes and, with the implementation of soft mobility hotspots, it would **reduce** the **number** of carbon vehicles that uses that roadway ans **replace** them for **bicycles**.

6.3 GLOBAL ACHIEVEMENTS FOR AOI

After describing the several achievements for the three parishes of our AoI, this section aims to summarise those output and results, in order to draw so conclusions regarding the detailed accomplishments.

As we can see in Table 7, the summarised information regarding the identification of new docking stations for *Marvila* and *Beato* is presented, but also a guideline for *Parque das Nações*, where there already fourteen docking stations implemented as stated previously.

Table 7: SSE - Sum of Squared Error for best K, following *Elbow method*

Council Parish Name	Number of Clusters	SSE - Sum of Squared Errors
Parque das Nações	14	0.853283
Marvila	8	1.662275
Beato	5	0.462912

Keeping in mind the suggested values for the docking stations implementation in our AoI, and after the detailed study regarding the three parishes and the potential instalation of such docking stations, some conclusions can be drawn, such as:

- *Parque das Nações* : several concentration points of devices with no docking stations (Area 3, in Section 6.1);
- More than 70% (10 out of 14) of docking stations in Parque das Nações distances 250m or less from algorithm' output;
- The others 4 docking stations are **suggested** for **uncovered areas** in *Parque das Nações*;
- *Beato* and *Marvila* : all docking stations are located on bicycle paths or distances less than 200m from mobility and education PoI;
- As seen, main avenues such as *Avenida Infante Dom Henrique* is one of the critical roadways in our study area.

Therefore, the **application** of **clustering methods**, followed by the methodology described in Chapter 5 and all the data transformations applied, have revealed an **accurate approach** to **identify new docking stations** for **bicycle sharing systems**.

Considering the correlation between other mobility services and soft mobility solutions potential usage have been identified, and also considering high affluence of people in train and subway stations, this fact accentuate the ideology sustained by Agency (2019), of soft mobility solutions and in particular, bicycle, being a **"first/last transport mode"**.

In summary, the closeness from in site docking stations and our methodology allows the validation of new docking stations, as seen in *Beato and Marvila*, providing an important **kickoff** for the **urban mobility planning** issue in **Lisbon**, in particular. Nevertheless, and still being related to BSSs, **this study** can be useful for **other soft mobility solutions** to be considered in **these locations**, such as **scooters** (electrical or not), since these kind of solutions integrate the **soft mobility** concept.

Part IV

FINAL CONSIDERATIONS

7

CONCLUSION

7.1 SUMMARY OF DISSERTATION AND RESULTS

7.1.1 *The purpose*

Urban mobility represents an issue of extreme importance for climate change purposes alongside with United Nations sustainable goals. Hence, is important to use sustainable modes of transportation, and also increasing the efficiency of existent systems already developed.

The main objective of this dissertation was to **identify new docking stations** for bicycle-sharing systems in three different council parishes of *Lisbon* city: *Beato*, *Marvila* and *Parque das Nações*. In this latest council parish, and as in other regions of *Lisbon* city center, there is already a bicycle-sharing system operating, denominated ***Gira - Lisbon Bicycles***. In this sense, and considering the possible growth and optimisation of the current system, it was implemented a cluster based approach in order to discover the new docking stations to be installed in the mentioned area.

Moreover, the importance of the BSS in the context of facing carbon emissions and air pollution is underline by the concept of **soft mobility**. This concept includes carbon free modes of transportation, and for that reason, eco-friendly. In this sense, soft mobility is one of the ways to reduce the climate changes effects and consequently, the carbon emissions.

In specific, *Lisbon* presents an interesting case study regarding urban mobility planning and traffic optimisation, since this city has been recognised as an hotspot for traffic jams and congestions, as mentioned in Chapter 1.

7.1.2 *Lisbon - Characterisation of AoI*

For that reason, CEiiA and this dissertation in particular, aimed to understand the existing mobility solutions available in *Lisbon* and study the transport modes that are currently operating in the three mentioned council parishes. As an output from such research and documentation, an important deliverable called "Lisbon Mobility baseline" was produced, being part of its content included in Chapter 3. In such chapter, we can find a geographical characterisation of several transport modes available such as train, subway and bus stop. Apart from that study, other geographical investigation regarding other important PoI in *Lisbon* and in particular, in our study area, allowed to deeply understand the existing infrastructures in our AoI, analyse important factors from *Lisbon* topographic and demographic indexes. Hence, it was possible to get a more reliable representation of the geographical environment in *Lisbon* city, and consequently, improve the quality of conclusions and associations made during the discussion of the results. For this task, several free data sources were used, in particular some portals regarding *Lisbon* data, such as *Lisboa Aberta* or *dados.gov*.

7.1.3 *Visual Data Exploration*

After recognising and studying the area of interest, it was important to understand and detail the **mobility patterns** that are present every day in Portugal's capital. Therefore, it was pretended to **identify** the most common clustered areas (with higher concentration of devices) in the map, understand the **increase of devices** in those critical areas, **identify main roadways** and **avenues** where **traffic jams** frequently occurs, the main places for day/night time frames, etc. To complete that task, a Portuguese telecommunication operator have provided some GSM data to make this study happen. The received data contained several important restrictions due to the GDPR and other compliance and legal guidelines that are currently being applied in Europe.

For that reason, the provided data was aggregated in order to respect all the legal requirements and the operating guidelines. Nevertheless, it was possible to access to interesting information, such as **day** and **hour** of the **operation**, **speed** recorded on that operation, **number of aggregated devices** and also **geometry** component, which represents the vectorial data, containing the coordinates of a polygon (which is similar to a square), which delimits the region where such operation have been performed, using the emission of GPS signal.

The main objectives for this study consisted in understand several mobility patterns in each council parish, and then, predict and identify bicycle-sharing docking stations. For example, in *Parque das Nações* it was notorious the **high affluence** and **high concentration** of devices nearby the **train** and **subway** station *Gare do Oriente*, but also nearby the subway station in *Moscavide*. Also, in this council parish, we have alerted and described the major traffic jams that causes several congestion in *Avenida Dom João II* and also in the *Avenida Infante Dom Henrique*. In particular, this avenue in Lisbon is such important in our study because it crosses the all three parishes we are working on. This data analysis, mostly visually task, was very important to get in direct contact with data, draw some conclusions and create some graphical artifacts that can express the ideas and reasons behind such data points.

7.1.4 *New docking stations identification methodology*

In order to identify new docking stations in our AoI, several mathematical and vectorial operations were performed, as described in Chapter 5. The aggregation of devices and the geographical representation of the data using a grid base, based on S2 cells, was obviously constraints and limitations that would impact the results of applying directly any kind of clustering technique. Hence, in order to reduce the geographical complexity of using a Polygon, a centroid methodology was used. This step allowed us to have atomic data points, i.e., represented by a single point (centroid), and so easily represent them in real-world maps, instead of "grid map cells" (polygon), which was the previous representation used. However, an additional limitation was still present: aggregated devices. In order to not be possible to identify single users or identify patterns, all samples were previously anonymised and aggregated. So, in order to keep the geographical balance for every data point, but at the same time keep the real-wold distribution regarding number of devices, an **unlist operation** was performed: **append/add** the same sample according to the N devices considered in the original sample. It was a **key step** in order to keep equitably the geospatial distribution of devices (data points). Combining such information with shapefiles of each council parish and applying intersection of each parish with data points for the mentioned parish, performed three different datasets, one for each parish.

After the preparation process, it was now possible to apply the **clustering algorithm** to the different datasets. The chosen algorithm was **K-Means** for several reasons. The first one was to **directly** compare with the state of the art of the current operation bicycle sharing system available in part of our AoI - *GIRA Bicycles* in *Parque das Nações*. Since K-Means Clustering is a single argument algorithm, namely the **number of clusters** that we pretend data to be clustered into, it granted the possibility to compare our clusters with the on-site stations for *Parque das Nações*. Due to this implementation, we were able to measure the distance between our suggestions for docking stations (algorithm output) and the existing docking stations. With such comparison, we had a **proof** of **quality** regarding all the **preparation data process**, and also, highlighting the **possibility** to **model** and **predict** soft mobility docking stations using **GSM data**. Nevertheless, choosing the K-Means algorithm had some constraint when applied to *Beato* and *Marvila*, where there are no bicycle sharing system operating (so far). In this case, we have used a well-known methodology to identify **suitable number of clusters** when this information is not set in the advance. This is called **Elblow method**, and as we studied in Chapter 2, let us to suggest a **suitable** number of clusters, according to the number and dispersion of the dataset.

Lastly, considering the whole work plan for this dissertation, an optimisation process was applied. Verified the importance of **closeness** to **bicycle lanes** and other **safe** paths do use bicycles and other soft mobility solutions, we have created a simple optimisation process: for each cluster output from the algorithm, it was **calculated** the **distance** to the **closet segment of bicycle lane**. If this distance was under 300 meters, then the docking stations should be placed over the bicycle lane. This will permit the potential/current users of such BSS to have some features such as **safety** and **security** maximised, since they can pick/leave any vehicle over the bicycle lane. As stated in several articles and papers, as seen in Chapter 2, reduces the risk of accident and, most important, fulfills one of the main requirements of soft mobility users: appropriate infrastructures for these systems.

7.1.5 *Insights from Results*

In this context, for *Parque das Nações* is suggested the addition of **four** new docking stations (essentially to the uncovered area of the current system), in *Marvila* install **seven new docking stations** and in *Beato* **five**. Considering mobility and urban planning real-world problem, it is also mentioned that the suggested numbers should be faced as the **minimum docking stations** since this kind of systems need **high geographical coverage** and also **closeness** between the docking stations points.

The truth is that several docking stations have been identified over this avenue, considering the high amount of vehicles and devices detected there. For *Marvila* and *Beato*, it was clear the correlation (once again) between the subway and train stations operating in this regions and the higher concentrations of devices. Nevertheless, considering the characteristics of such council parish, where the major concentrations of devices occurs in residential building during the night, it was possible to identify such clusters and suggest/predict such concentration areas as suitable for installation of new docking stations. Also, *Avenida Infante Dom Henrique* has revealed more traffic jams in *Marvila and Beato's* extension, and also there, docking stations were identified as described in Section 6.

Firstly, with regard to *Parque das Nações*, a direct comparison with the existing docking stations operated by the *Gira - Lisbon Bicycles* system, we have found out a similar configuration for the docking stations in Area 1 and Area 2, as described in Chapter 6. Nevertheless, and despite having found out more than 70% of docking stations in *Parque das Nações* distancing less than 300 meters from the existing ones, it was discovered the need of adding four new docking stations. This four docking stations, located in the dominated Area 3 in the previous Chapter, represent as **uncovered area** in the existing system, which means an identification of **potential expansion** of the current system, verified by our results.

Briefly, the proposed docking stations findings suggest the addition of, at least, five docking stations in *Beato*. This council parish, which is the smallest one among our AoI, has 2,46 km^2 area and as seen documented before, has the lower concentration of devices among the three council parishes considered in this study. This docking stations are set to serve around 10 thousand inhabitants. Regarding *Marvila*, this is the biggest council parish concerning area (7,12 km^2), but our analysis have demonstrated a lower concentration of devices (when compared to *Parque das Nações*), and for that reason, it was identified eight docking stations as minimum number of docking stations to be installed. In both council parishes, the docking stations identified have been proposed close to several interesting PoI such as schools, train/subway stations or nearby bicycle lanes.

7.1.6 *Limitations and constraints*

Regarding the development of this project, and considering all the process behind its implementation, it is necessary to emphasise the **key role** of the **data** used. Noticing that our devices were represented by the GSM data provided by the telecommunication operator, that highlights the **sensitivity** of such data.

Thus, and as mentioned before, the GDPR had an important role during this research, making necessary to work with **aggregated data**. This aggregations represent additional error to the final results obtained, since it not possible to work with the exact location of the devices, due to privacy guidelines.

Moreover, and in addition to the geographical restrictions mentioned, temporal constraints were also faced. In this case, the provided GSM data only contemplates a time frame for January 2020, being also an important factor to point out, since this is a short time window. For that reason, and considering weather plays a key role in mobility choices, this results are related to the time window mentioned.

In a simple description, the main limitations and constraints faced are related to the GPDR and legal guidelines that must be followed. Nevertheless, the obtained results are truly indicative of an interesting approach to identify the location of docking stations.

7.1.7 *The system and its applicability*

The development of the described decision support system may help the competent authorities, in this case *Lisbon Municipality*, to better planing the geographical distribution of the available bike sharing system. However, the importance of that study and the findings that were shared, reveal a potential generalisation of the process and consider, in the future, the installation of other soft-mobility solutions such as scooters of roller-skating in the identified locations, since it can serve the population in different mobility choices but keeping the identified locations that comes from this study.

Considering the explained high level approach used for Lisbon, and being this a data driven methodology, it means that the described **decision support system** can be replicated in other cities all over the world, since some kind of geographical representation of devices/people can be provided.

Making a reflection regarding the work scheduled and presented in Section 1.4, it is now possible to verify that the **scheduled plan** was almost completed in time, with little variations due to external factors such as data providing process and release, being the dissertation writing process the one who took a little bit more time than initially expected. Overall, the scheduled and agreed technical work plan was completed on time, being some of the adjustments justified with some extra activities, described in more detail in Section 7.4.

In conclusion, let us say that the **main objective** of **suggesting** and **predicting** new bike sharing docking stations in Lisbon have been completed with **clear success**, finding a new approach to accurately identify cluster hotspots of devices (i.e., people), based on mathematical and vectorial processes over GSM data, which have been revealed a novelty to the scientific community. This data-driven approach is now a suitable kickoff for the installation in the future of new docking stations in Lisbon, but also for the **bicycle lanes** extension program that *Lisbon Municipality* have already shared the intention to work on in the upcoming months/years.

7.2 FUTURE WORK

Regarding the results that were found with the realisation of this dissertation, it is important to remember some constraints that were faced during this process, in order to potentially increase the accuracy of our results.

As noted in Chapter 4, an initial constraint was considered: the temporal distribution of our data. Unfortunately, the data provided by the telecommunication operator was for January of 2020, and of course, the results obtained cannot be completely generalised as if the whole year was considered. For that reason, and considering the insights that were drawn from data exploration process and also the findings regarding the docking stations installation, it should be possible, in partnership with CEiiA, develop sensors that can capture the amount of devices that crosses their covered area, for example, or even capture a larger amount of data, during several weeks/months. This would allow us to better understand the devices distribution in a wider timeline, which is very important so that we can notice the influence of other external factors in mobility context such as the weather conditions, holidays, etc.

In addition, as future improvements and considering a research path, **new clustering algorithms** could be applied and tested within the current methodology. This will let us compare the existing results and evaluation metrics defined, in order to possibly improve the current achievements we have reached. For example, one of this algorithms is **K-Medoids**. As seen in State of the Art (Chapter 2), this algorithm provides a similar behaviour when compared to the **K-Means Clustering**. Nevertheless, it differs in the **centroids calculation process**, forcing that the **identified cluster** must **belong** to the **initial dataset**. According to some authors, this algorithm still lack of optimisation and for that reason presents lower rate of usage in **high volume** of **data problems**, and also, lower performance when directly compared with K-Means.

Moreover, supplementary analysis to the clusters output from our system could be done, namely with **bus stops points**. Considering the macro analysis with the docking stations locations with other PoI such as subway stations, train stations, schools etc, it would also be interesting to apply a micro analysis with more detail, for example, with bus stops or train stations would be also interesting, since it would allow to draw even deeper conclusions regarding the soft mobility relation with other public transport modes. In fact, soft mobility vehicles are commonly seen in literature as "first/last mile modes", since their usage is maximised when combined with other existing solutions, unlike being a direct alternative to person car, motorbike, etc. This is justified, in some cases, due to other external factors that affect mobility choices such as weather, length of journeys, etc.

7.3 SCIENTIFIC CONTRIBUTIONS

As direct output from this master dissertation research project, the Fontes et al. (2022) paper have been produced. This paper aims to summarise the information presented in this dissertation, aiming to share with scientific community the advances of this investigation, in order to **highlight** the **usefulness** of **AI algorithms** to solve real world problems, in this case, **mobility planning problems** taking advantage from **clustering algorithms**. Additionally, it pretends to give relevance to the fact that possible to work with **anonymized data**, with **several restrictions** due to the GDPR, but also achieve interesting results in the mobility context, so that it can be possible to reduce the negative side effects of mobility.

Moreover, the paper has been accepted in the 18th International Conference of Distributed Computing and Artificial Intelligence (DCAI), which took place in *Salamanca*, Spain on October, 2021. The participation in such event, further in-person, provided an amazing experience to share knowledge, contact with other researchers, enthusiasts and professionals from AI sphere all over the world. This sharing time helped me to develop capabilities and other important skills such as communication, presentation, foreign language, among other important influences that only the in-site participation can provide.

For all the mentioned reasons and beyond, the chosen path of writing and publishing a paper in this dissertation context is a proof of such important study, recognising the quality of everyone involved in the project, alongside the developed work and results obtained.

Moreover, and after the presentation and publication of the mentioned paper, the group of authors have been invited to submit a **extended improved** version of the current paper, with new naming and with no page limitation, in order to provide a deeper view of these findings. This invitation for new publication is set to be integrated in a Special Issue named "Challenges for the Development of Sustainable Smart Cities”, which is an international, scientific, peer-reviewed, open access journal on the science and technology of smart cities, published quarterly online by the Multidisciplinary Digital Publishing Institute (MDPI), and accessible in MDPI (2021).

7.4 CEIIA EXPERIENCE

At CEiiA, I have grown up in so many aspects that, this dissertation experience, proofs how much someone can improve in several areas in a short period. First of all, the pleasure of integrating a Data Science team was a very important step in my academic formation, since it provided a "long term" experience with professional environment and people, which allowed me to rapidly be even more rigorous (in the good sense, of course) work plan and state of mind. The continuous interaction with every member of the team, let me develop communication and other soft skills that, nowadays, are so important as the technical component of any project.

Regarding technical participation, I had the opportunity to collaborate on many projects that CEiiA is currently working on. Hence, I had the pleasure to work alongside the team in two distinct areas: **precision agriculture** and **satellite image** processing with AI, namely vessel detection.

Regarding precision agriculture, CEiiA is working alongside other companies to take advantage of AI solutions to solve some precision agriculture issues, where images from fields and plantations are collected with Unmanned Aerial Vehicle (UAV). From these images, it is pretended to detect and identify crops of plantations from other data (such as roads, buildings, etc).

Concerning the vessel detection context, CEiiA is collaborating with European Space Agency(ESA) to detect vessels from satellite images, since it is pretended to act in illegal activity detection, for example. For that reason, the participation and working on these projects allowed me to discuss and learn from AI solutions and how to work with such GIS data to solve, once again, real-world problems.

Also, in the background there this dissertation is inserted, CEiiA is working on other important use cases for urban mobility planning and solutions in Lisbon city, which I have had the opportunity to join. Thus, CEiiA gave me the please to be part of the brainstorming meetings with some partners and, obviously, inside the team and the department I was involved in, meaning a lot to me, since this is a key signal of confidence and responsibility.

Once again, I must appreciate everyone who guided and worked with me, keeping the support I needed and pushing me to keep working and staying motivated. This dissertation, in some way, represents everyone that have shared important moments with me, and of course, have given me important inputs. I will be forever in debt.

BIBLIOGRAPHY

Abdul Qureshi. Managing data science projects using crisp-dm process framework, 2019. https://medium.com/@aqureshi/managing-data-science-projects-using-crisp-dm-, Last accessed on 29-01-2021.

European Environment Agency. The first and last mile — the key to sustainable urban transport transport and environment report 2019, 2019. https://www.eea.europa.eu/publications/the-first-and-last-mile/, Last accessed on 29-10-2021.

European Environment Agency. Air pollution is the biggest environmental health risk in europe, 2021. https://www.eea.europa.eu/themes/air, Last accessed on 30-09-2021.

AIM . Top 10 best geospatial tools for gis, 2019. https://analyticsindiamag.com/top-10-best-geospatial-tools-gis-mapping-data-visualization/, Last accessed on 28-04-2021.

Akker, van den. Ai mobility, 2021. https://www.uu.nl/en/research/human-centered-artificial-intelligence/special-interest-groups/ai-mobility, Last accessed on 16-11-2021.

Vito Albino, Umberto Berardi, and Rosa Maria Dangelico. Smart cities: definitions, dimensions, performance, and initiatives. *Journal of Urban Technology*, 22(1):3–21, January 2015. ISSN 1063-0732, 1466-1853. doi: 10.1080/10630732.2014.942092. URL http://www.tandfonline.com/doi/full/10.1080/10630732.2014.942092.

Saima Bano and Naeem Khan. A survey of data clustering methods. *International Journal of Advanced Science and Technology*, 113, 04 2018. doi: 10.14257/ijast.2018.113.14.

Marcelo Beckmann, Nelson Ebecken, and Beatriz Lima. A knn undersampling approach for data balancing. *Journal of Intelligent Learning Systems and Applications*, 2015.

Mafalda Bernardo, Manuela Aparicio, and Miguel Neto. Smart mobility: a multimodal services study in the metropolitan area of lisbon. 10 2019.

CEiiA. Ceiia - centre of engineering and product development. https://www.ceiia.com/, Last accessed on 15-10-2021.

CEiiA. Ceiia participa na cop26, reforÇando o seu papel em ser um ator ativo na descarbonizaÇÃo das cidades, 2021. https://www.ceiia.com/single-post/ceiiacop26, Last accessed on 15-10-2021.

Check Point. Gsm overview, 2021. https://sc1.checkpoint.com/documents/R81/WebAdminGuides/EN/CP_R81_CarrierSecurity_AdminGuide/Topics-CSG/GSM-Overview.htm/, Last accessed on 05-07-2021.

Federico Chiariotti, Chiara Pielli, Andrea Zanella, and Michele Zorzi. A dynamic approach to rebalancing bike-sharing systems. *Sensors*, 18(2):512, February 2018. ISSN 1424-8220. doi: 10.3390/s18020512. URL http://www.mdpi.com/1424-8220/18/2/512.

Christina Voskoglou. What is the best programming language for machine learning?, 2017. https://towardsdatascience.com/what-is-the-best-programming-language-for-machine-learning, Last accessed on 07-02-2021.

Climate ADAPT. National circumstances relevant to adaptation actions, 2021. https://climate-adapt.eea.europa.eu/countries-regions/countries/portugal, Last accessed on 02-02-2021.

Climate Change Conference. Cop26 goals, 2021. https://ukcop26.org/cop26-goals/, Last accessed on 17-10-2021.

CML. *Diagnístico Social de Lisboa - Retrato das Freguesias*, 2011. https://www.am-lisboa.pt/documentos/1532873203W6wPC8in0Aj47ZK8.pdf, Last accessed on 20-05-2021.

European Commission. European climate pact, 2020. URL https://ec.europa.eu/clima/eu-action/european-green-deal/european-climate-pact_pt. [Access at 28-01-2021].

Yash Dagli. Partitional clustering using clarans method, 2019. https://medium.com/analytics-vidhya/partitional-clustering-using-clarans-method, Last accessed on 29-06-2021.

Mapa de Lisboa. *Mapas e informação turística de Lisboa*, 2021. https://www.mapalisboa.net/, Last accessed on 13-05-2021.

de Vries, Andrie , Meys, Joris. The benefits of using r. https://www.dummies.com/programming/r/the-benefits-of-using-r/, Last accessed on 13-03-2021.

Paul DeMaio. Bike-sharing: history, impacts, models of provision, and future. *Journal of Public Transportation*, 12(4):41–56, December 2009. ISSN 1077-291X, 2375-0901. doi: 10.5038/2375-0901.12.4.3. URL http://scholarcommons.usf.edu/jpt/vol12/iss4/3/.

Direção Geral do Território Portugal. *Carta Administrativa Oficial de Portugal*, 2020. https://www.dgterritorio.gov.pt/cartografia/cartografia-tematica/caopo, Last accessed on 25-05-2021.

EMEL. Bicicletas gira já rolaram 1 milhão de viagens, 2018. https://www.emel.pt/pt/noticias/bicicletas-gira-ja-rolaram-1-milhao-de-viagens/, Last accessed on 07-05-2021.

EMEL. Lisbon bicycles 2018 statistics, 2018. https://www.emel.pt/pt/noticias/bicicletas-gira-ja-rolaram-1-milhao-de-viagens-2-2/, Last accessed on 16-05-2021.

European Commission. Sustainable urban mobility must come first, 2020. https://cor.europa.eu/pt/news/Pages/sustainble-urban-mobility-must-come-first.aspx, Last accessed on 19-10-2021.

European Union. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016, 2016. https://eur-lex.europa.eu/eli/reg/2016/679/oj, Last accessed on 10-02-2021.

B Everitt, Sabine Landau, M. Leese, and Daniel Stahl. *Cluster Analysis*. 01 2011. ISBN 9780470749913. doi: 10.1002/9780470977811.ch8.

Felix, Rosa and Moura, Filipe. Socio-economic externalities of bike sharing in lisbon. 2020. https://tecnico.ulisboa.pt/en/news/study-carried-out-at-tecnico-highlighted-by-national-geographic/, Last accessed on 10-05-2021.

Tiago Fontes, Miguel Arantes, P. V. Figueiredo, and Paulo Novais. Bike-sharing docking stations identification using clustering methods in lisbon city. In Kenji Matsui, Sigeru Omatu, Tan Yigitcanlar, and Sara Rodríguez González, editors, *Distributed Computing and Artificial Intelligence, Volume 1: 18th International Conference*, pages 200–209, Cham, 2022. Springer International Publishing. ISBN 978-3-030-86261-9.

Brendan J. Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, 315(5814): 972–976, 2007. doi: 10.1126/science.1136800. URL https://www.science.org/doi/abs/10.1126/science.1136800.

Rosa Félix, Filipe Moura, and Kelly J. Clifton. Maturing urban cycling: Comparing barriers and motivators to bicycle of cyclists and non-cyclists in lisbon, portugal. *Journal of Transport Health*, 15:100628, 2019. ISSN 2214-1405. doi: https://doi.org/10.1016/j.jth.2019.100628. URL https://www.sciencedirect.com/science/article/pii/S2214140518306054.

Gopi Gandhi and Rohit Srivastava. Review paper: A comparative study on partitioning techniques of clustering algorithms. *International Journal of Computer Applications*, 87:10–13, 2014.

Geodados - CML. Plataforma de dados abertos georreferenciados da câmara municipal de lisboa, 2021. https://geodados-cml.hub.arcgis.com/, Last accessed on 05-06-2021.

Paolo Giordani, Maria Ferraro, and Francesca Martella. *Hierarchical Clustering*, pages 9–73. 08 2020. ISBN 978-981-13-0552-8. doi: 10.1007/978-981-13-0553-5_2.

GIRA. Mapa estações gira, 2021. https://www.gira-bicicletasdelisboa.pt/descobre-as-estacoes/, Last accessed on 14-05-2021.

Joonho Gong and Hyunjoong Kim. RHSBoost: Improving classification performance in imbalance data. *Computational Statistics & Data Analysis*, 111:1–13, July 2017. ISSN 01679473. doi: 10.1016/j.csda.2017.01.005. URL https://linkinghub.elsevier.com/retrieve/pii/S016794731730018X.

GPS.gov. Gps accuracy and performance, 2021. https://www.gps.gov/systems/gps/performance/accuracy/, Last accessed on 03-02-2021.

GrindGIS. Pros and cons of qgis, 2019. https://grindgis.com/software/pros-and-cons-of-qgis, Last accessed on 10-03-2021.

Mark Haakman, Luís Cruz, Hennie Huijgens, and Arie Deursen. Ai lifecycle models need to be revised. an exploratory study in fintech. 10 2020.

Michael Haenlein and Andreas Kaplan. A brief history of artificial intelligence: on the past, present, and future of artificial intelligence. *California Management Review*, 61(4):5–14, August 2019. ISSN 0008-1256, 2162-8564. doi: 10.1177/0008125619864925. URL http://journals.sagepub.com/doi/10.1177/0008125619864925.

Subhan Hartanto, Mhd Furqan, Andysah Putera Utama Siahaan, and Wirda Fitriani. Haversine method in looking for the nearest masjid. *International Journal of Engineering Research*, 3:187–195, 08 2017. doi: 10.23883/IJRTER.2017.3402.PD61H.

IAV. Artificial intelligence for better mobility, 2021. https://www.iav.com/en/what-moves-us/artificial-intelligence-for-better-mobility/, Last accessed on 25-10-2021.

IMT. Ciclando - plano nacional de promoção da bicicleta e outros modos suaves. 2013-2020.

INE. Census 2011, 2011. https://censos.ine.pt/xportal/xmain?xpid=INE&xpgid=censos2011_apresentacao, Last accessed on 20-02-2021.

INE - Instituto Nacional Estatística. Pessoas em movimento : estatísticas sobre a mobilidade na europa, 2019. https://ine.pt/scripts/EuMove_2019/bloc-3a.html?lang=pt, Last accessed on 10-02-2021.

European Commission INE. The foreign population in portugal 2011, 2011. https://ec.europa.eu/migrant-integration/library-document/foreign-population-portugal-2011_en, Last accessed on 28-05-2021.

Jan Wisniewski - RESETI. Paris's public bike sharing schemes: Win or fail for the environment?, 2019. https://en.reset.org/blog/pariss-public-bike-sharing-schemes-win-or-fail-environment-09302019, Last accessed on 05-02-2021.

Xin Jin and Jiawei Han. *K-Medoids Clustering*, pages 564–565. Springer US, Boston, MA, 2010. ISBN 978-0-387-30164-8. doi: 10.1007/978-0-387-30164-8_426. URL https://doi.org/10.1007/978-0-387-30164-8_426.

Xin Jin and Jiawei Han. Partitional clustering. 01 2011. doi: 10.1007/978-0-387-30164-8_631.

Leonard Kaufman and Peter Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. 09 2009. ISBN 9780470317488.

Koleva, Nancy . Ai and data science lifecycle: Key steps and considerations, 2019. https://blog.dataiku.com/ai-projects-lifecycle-key-steps-and-considerations, Last accessed on 28-04-2021.

Sören Krach, Frank Hegel, Britta Wrede, Gerhard Sagerer, Ferdinand Binkofski, and Tilo Kircher. Can machines think? interaction and perspective taking with robots investigated via fmri. *PloS one*, 3:e2597, 07 2008. doi: 10.1371/journal.pone.0002597.

Satyam Kumar. Understanding k-means, k-means++ and, k-medoids clustering algorithms, 2020. https://towardsdatascience.com/understanding-k-means-k-means-and-k-medoids-clustering-algorithms, Last accessed on 21-06-2021.

Kumar, Vikash . Python vs r: What's best for machine learning?, 2019. https://towardsdatascience.com/python-vs-r-whats-best-for-machine-learning, Last accessed on 07-02-2021.

Rosa Anna La Rocca. Soft Mobility and Urban Transformation. *TeMA - Journal of Land Use, Mobility and Environment 2*, 2010. URL https://www.researchgate.net/publication/279669946_Soft_Mobility_and_Urban_Transformation.

Debra Lam and Peter Head. Sustainable urban mobility. In Oliver Inderwildi and Sir David King, editors, *Energy, Transport, & the Environment*, pages 359–371. Springer London, London, 2012. ISBN 9781447127161 9781447127178. doi: 10.1007/978-1-4471-2717-8_19. URL http://link.springer.com/10.1007/978-1-4471-2717-8_19.

Dongheui Lee, Dana Kulic, and Yoshihiko Nakamura. Missing motion data recovery using factorial hidden Markov models. Pasadena, California, USA, 2008. URL https://www.researchgate.net/publication/221070411_Missing_motion_data_recovery_using_factorial_hidden_Markov_models.

Myeong-Woo Lee, Grant McKenzie, and Rajat Aghi. Exploratory cluster analysis of urban mobility patterns to identify neighborhood boundaries. 2017.

Carlos Lemonde, Elisabete Arsénio, and Rui Henriques. Exploring multimodal mobility patterns with big data in the city of lisbon. 2020.

Luís Liberato. O potencial dos sistemas de bicicletas partilhadas: uma contribução para a construção do paronama português. Master's thesis, Faculdade de Engenharia - Universidade do Porto, january 2018.

Aristidis Likas, Nikos Vlassis, and Jakob J. Verbeek. The global k-means clustering algorithm. *Pattern Recognition*, 36(2):451–461, 2003. ISSN 0031-3203. doi: https://doi.org/10.1016/S0031-3203(02)00060-2. URL https://www.sciencedirect.com/science/article/pii/S0031320302000602. Biometrics.

Lisboa Aberta - CML. Lisboa aberta, 2021. http://lisboaaberta.cm-lisboa.pt/index.php/pt/, Last accessed on 05-06-2021.

Maayan, Gilad. How ai is changing the mobility landscape, 2020. https://www.dataversity.net/how-ai-is-changing-the-mobility-landscape/, Last accessed on 23-10-2021.

Machado, Pedro. Sump and shared mobility in lisbon: is soft regulation sustainable?, 2019. https://www.eltis.org/sites/default/files/sump2019_c1_machado_pedro_lisbon.pdf, Last accessed on 03-05-2021.

Elżbieta Macioszek, Paulina Świerk, and Agata Kurek. The bike-sharing system as an element of enhancing sustainable mobility—a case study based on a city in poland. *Sustainability*, 12(8), 2020. ISSN 2071-1050. doi: 10.3390/su12083285. URL https://www.mdpi.com/2071-1050/12/8/3285.

Luis M. Martinez, Luís Caetano, Tomás Eiró, and Francisco Cruz. An optimisation algorithm to establish the location of stations of a mixed fleet biking system: An application to the city of lisbon. *Procedia - Social and Behavioral Sciences*, 54:513–524, 2012. ISSN 1877-0428. doi: https://doi.org/10.1016/j.sbspro.2012.09.769. URL https://www.sciencedirect.com/science/article/pii/S1877042812042310. Proceedings of EWGT2012 - 15th Meeting of the EURO Working Group on Transportation, September 2012, Paris.

Mary K. Pratt - TechTarget. Ict (information and communications technology, or technologies), 2019. https://searchcio.techtarget.com/definition/ICT-information-and-communications-technology-or-technologies/, Last accessed on 08-02-2021.

Isabel Matias and Ana Virtudes. Cycling mobility in slopping cities: Trondheim and other lessons. *KnE Engineering*, 5(5):139–151, May 2020. doi: 10.18502/keg.v5i5.6931. URL https://knepublishing.com/index.php/KnE-Engineering/article/view/6931.

John McCarthy and Edward A. Feigenbaum. In memoriam: Arthur samuel: Pioneer in machine learning. *AI Magazine*, 11(3):10, Sep. 1990. doi: 10.1609/aimag.v11i3.840. URL https://ojs.aaai.org/index.php/aimagazine/article/view/840.

MDPI. Smart cities, 2021. https://www.mdpi.com/journal/smartcities, Last accessed on 14-12-2021.

Michael Majster et al. Artificial intelligence in mobility, 2021. https://www.adlittle.com/en/insights/viewpoints/artificial-intelligence-mobility, Last accessed on 23-10-2021.

Peter Midgley. BICYCLE-SHARING SCHEMES: ENHANCING SUSTAINABLE MOBILITY IN URBAN AREAS. 2011. URL https://static.un.org/esa/dsd/resources/res_pdfs/csd-19/Background-Paper8-P.Midgley-Bicycle.pdf.

Mike Thurber. A holistic framework for managing data analytics projects, 2020. https://www.elderresearch.com/blog/a-holistic-framework-for-managing-data-analytics-projects/, Last accessed on 29-01-2021.

Motor24. The planned 200 km of bicycle planes in lisbon, 2019. https://www.motor24.pt/sites/wattson/estamos-a-dois-anos-de-termos-200-km-de-ciclovias-em-lisboa/542831/, Last accessed on 22-09-2021.

NACTO. High-quality bike facilities increase ridership and make biking safer, 2016. https://nacto.org/2016/07/20/high-quality-bike-facilities-increase-ridership-make-biking-safer/, Last accessed on 30-05-2021.

nationalgrid. Cop26 – coming together to tackle climate change, 2021. https://www.nationalgrid.com/responsibility/environment/cop26, Last accessed on 17-10-2021.

United Nations. 17 sustainable development goals, 2015. URL https://sdgs.un.org/goals. [Access at 29-01-2021].

The Portugal News. Lisbon is most congested city in the iberian peninsula, 2017. https://www.theportugalnews.com/news/lisbon-is-most-congested-city-in-the-iberian-peninsula/41154, Last accessed on 27-10-2021.

Jordi Nofre and Jorge Sequera. Sequera, j., nofre, j. (2020). touristification, transnational gentrification and urban change in lisbon: The neighbourhood of alfama. urban studies, 57(15), 3169-3189.. doi: 10.1177/0042098019883734. *Urban Studies*, 57:3169–3189, 11 2020. doi: 10.1177/0042098019883734.

Observador. More 700 bicycles in lisbon until march 2021, 2021a. https://observador.pt/2020/12/30/lisboa-commais-700-bicicletas-eletricas-final-marco-2021/, Last accessed on 21-05-2021.

Observador. Alteração ciclovias em lisboa, 2021b. https://observador.pt/2021/11/02/carlos-moedas-diz-que-vai-estudar-todas-as-ciclovias-de-lisboa-, Last accessed on 10-11-2021.

Observatorio da Emigracao. Emigração portuguesa 2020, 2020. http://observatorioemigracao.pt/np4/7785.html, Last accessed on 29-05-2021.

OpenStreetMap. Open street map, 2021. https://www.openstreetmap.org/#map=7/39.602/-7.839, Last accessed on 29-09-2021.

Ashish Patel. Machine learning algorithm overview, 2018. https://medium.com/ml-research-lab/machine-learning-algorithm-overview-5816a2e6303, Last accessed on 27-03-2021.

Portal Dados Abertos Administração Pública. Portal dados.gov.pt, 2021. https://dados.gov.pt/pt/, Last accessed on 04-06-2021.

Portuguese American Journal. Lisbon: Bike sharing service available for locals and visitors – portugal, 2021. https://portuguese-american-journal.com/lisbon-bike-sharing-service-available, Last accessed on 02-02-2021.

Priyanki Baruah . 20 differences between qgis and arcgis, 2019. https://planningtank.com/geographic-information-system/differences-qgis-arcgis, Last accessed on 25-04-2021.

Python Org. Python - programming language, 2021. https://www.python.org/, Last accessed on 25-10-2021.

QGIS . Qgis documentation, 2021. https://www.qgis.org/en/site/index.html, Last accessed on 25-04-2021.

R. The r project for statistical computing, 2021. https://www.r-project.org/, Last accessed on 29-03-2021.

Sabado. Increase number of lisbon bicycles, 2021. https://www.sabado.pt/portugal/detalhe/emelestima-duplicar-numero-de-bicicletas-gira-2021, Last accessed on 20-05-2021.

Shagan Sah. Machine learning: A review of learning types. 07 2020. doi: 10.20944/preprints202007.0230.v1.

Joerg Sander. *Density-Based Clustering*, pages 270–273. Springer US, Boston, MA, 2010. ISBN 978-0-387-30164-8. doi: 10.1007/978-0-387-30164-8_211. URL https://doi.org/10.1007/978-0-387-30164-8_211.

Alberto Sanmiguel-Rodríguez and Víctor Arufe-Giráldez. Impact of climate on a bike-sharing system. minutes of use depending on day of the week, month and season of the year. *Cuadernos de Psicologia del Deporte*, 19: 102–112, 05 2019. doi: 10.6018/cpd.338441.

Pulkit Sharma. Comprehensive guide to k-means clustering, 2019. https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/, Last accessed on 17-06-2021.

Sheromova, Vasilisa. A brief history of python, 2020. https://exyte.com/blog/a-brief-history-of-python, Last accessed on 20-5-2021.

Nikita Shiledarbaxi. Comprehensive guide to clarans clustering algorithm, 2021. https://analyticsindiamag.com/comprehensive-guide-to-clarans-clustering-algorithm/, Last accessed on 28-06-2021.

Rajeev Pratap Singh, Anita Singh, and Vaibhav Srivastava, editors. *Environmental issues surrounding human overpopulation*. Advances in environmental engineering and green technologies. Information Science Reference, Hershey, 2017. ISBN 9781522516835 9781522516842.

Adish Singla, Marco Santoni, Gÿbor Bartók, Pratik Mukerji, Moritz Meenen, and Andreas Krause. Incentivizing users for balancing bike sharing systems. January 2015. URL https://dl.acm.org/doi/10.5555/2887007.2887108.

Smar50Awards. *Lisboa distinguida como Smart City*, 2019. https://www.smartplanet.pt/news/smart-cities/lisboa-distinguida-como-smart-city, Last accessed on 07-02-2021.

Zanfina Svirca. Density-based algorithms, 2020. https://towardsdatascience.com/density-based-algorithms-49237773c73b, Last accessed on 30-08-2021.

G. Sánchez-Barroso, J. González-Domínguez, J. García-Sanz-Calcedo, and M. Sokol. Impact of weather-influenced urban mobility on carbon footprint of spanish healthcare centres. *Journal of Transport Health*, 20: 101017, 2021. ISSN 2214-1405. doi: https://doi.org/10.1016/j.jth.2021.101017. URL https://www.sciencedirect.com/science/article/pii/S2214140521000116.

Karolina Taczanowska, A. Muhar, and C. Brandenburg. Potential and limitations of gps tracking for monitoring spatial and temporal aspects of visitor behaviour inrecreational areas. *Management for protection and sustainable development*, pages 451–455, 10 2008.

Akbar Telikani, Amirhessam Tahmassebi, Wolfgang Banzhaf, and Amir H. Gandomi. Evolutionary machine learning: A survey. *ACM Comput. Surv.*, 54(8), oct 2021. ISSN 0360-0300. doi: 10.1145/3467477. URL https://doi.org/10.1145/3467477.

Câmara Municipal de Lisboa. *Economia de Lisboa em Números - 2020*, 2020. https://issuu.com/camara_municipal_lisboa/docs/economia_lisboa_em_numeros_2020, Last accessed on 13-05-2021.

Velmurugan Thambusamy and Santhanam T. A survey of partition based clustering algorithms in data mining: An experimental approach. *Information Technology Journal*, 10, 03 2011. doi: 10.3923/itj.2011.478.484.

TopographicMap. Elevation map of lisbon, 2021. https://en-gb.topographic-map.com/maps/wvd2/Lisbon/, Last accessed on 21-05-2021.

Alan M. Turing. *Computing Machinery and Intelligence*, pages 23–65. Springer Netherlands, Dordrecht, 2009. ISBN 978-1-4020-6710-5. doi: 10.1007/978-1-4020-6710-5_3. URL https://doi.org/10.1007/978-1-4020-6710-5_3.

Yannis Tyrinopoulos and Constantinos Antoniou. Factors affecting modal choice in urban mobility. *European Transport Research Review*, 5:27–39, 03 2013. doi: 10.1007/s12544-012-0088-3.

United Nations. 2030 agenda for sustainable development, 2015. https://sdgs.un.org/2030agenda, Last accessed on 27-01-2021.

UShift. Lisbon street's inclinations, 2021. http://ushift.tecnico.ulisboa.pt/data/, Last accessed on 22-05-2021.

Hao Wang and Dit-Yan Yeung. Towards bayesian deep learning: a framework and some existing methods. *IEEE Transactions on Knowledge and Data Engineering*, 28(12):3395–3408, December 2016. ISSN 1041-4347. doi: 10.1109/TKDE.2016.2606428. URL http://ieeexplore.ieee.org/document/7562516/.

Rixin Wang, Xuebing Gong, Minqiang Xu, and Yuqing Li. Fault detection of flywheel system based on clustering and principal component analysis. *Chinese Journal of Aeronautics*, 28, 10 2015. doi: 10.1016/j.cja.2015.10.003.

Julia Winslow and Oksana Mont. Bicycle sharing: Sustainable value creation and institutionalisation strategies in barcelona. *Sustainability*, 11(3), 2019. ISSN 2071-1050. doi: 10.3390/su11030728. URL https://www.mdpi.com/2071-1050/11/3/728.

Peng Xie, Tianrui Li, Jia Liu, Shengdong Du, Xin Yang, and Junbo Zhang. Urban flow prediction from spatiotemporal data using machine learning: A survey. *Information Fusion*, 59:1–12, July 2020. ISSN 15662535. doi: 10.1016/j.inffus.2020.01.002. URL https://linkinghub.elsevier.com/retrieve/pii/S1566253519303094.

Dongkuan Xu and Yingjie Tian. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2, 08 2015. doi: 10.1007/s40745-015-0040-1.

Zidong Yang, Ji Hu, Yuanchao Shu, Peng Cheng, Jiming Chen, and Thomas Moscibroda. Mobility modeling and prediction in bike-sharing systems. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, pages 165–178, Singapore Singapore, June 2016. ACM. ISBN 9781450342698. doi: 10.1145/2906388.2906408. URL https://dl.acm.org/doi/10.1145/2906388.2906408.

Xiuwen Yi, Yu Zheng, Junbo Zhang, and Tianrui Li. ST-MVL: filling missing values in geo-sensory time series data. 2016. URL https://dl.acm.org/doi/10.5555/3060832.3060999.

Chunhui Yuan and Haitao Yang. Research on k-value selection method of k-means clustering algorithm. *J*, 2(2):226–235, 2019. ISSN 2571-8800. doi: 10.3390/j2020016. URL https://www.mdpi.com/2571-8800/2/2/16.

Mohamed Zayed. Towards an index of city readiness for cycling. *International Journal of Transportation Science and Technology*, 5, 01 2017. doi: 10.1016/j.ijtst.2017.01.002.

Junbo Zhang, Yu Zheng, Dekang Qi, Ruiyuan Li, and Xiuwen Yi. DNN-based prediction model for spatio-temporal data. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 1–4, Burlingame California, October 2016. ACM. ISBN 9781450345897. doi: 10.1145/2996913.2997016. URL https://dl.acm.org/doi/10.1145/2996913.2997016.