NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

An exploratory analysis of Pedestrian Accidents patterns in Lisbon

Francisco José Monteiro do Espírito Santo

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

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

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

AN EXPLORATORY ANALYSIS OF PEDESTRIAN ACCIDENTS PATTERNS IN LISBON

by

Francisco José Monteiro do Espírito Santo

Project Work report presented as partial requirement for obtaining the Master's degree in Information Management/ Master's degree in Statistics and Information Management, with a specialization in Knowledge Management and Business Intelligence

Co Advisor: Miguel de Castro Neto

Co Advisor*:* Pedro Sarmento

July 2021

ACKNOWLEDGEMENTS

The time has come for me to express my deep and sincere acknowledgements to everyone!

I would like to thank Lisboa Aberta team, for sharing the data, as well as for all the help given throughout the project.

Special thanks to Professor Miguel Castro Neto and researcher Pedro Sarmento. Thank you for your advices and constant feedback.

To my family and friend, for all the patience, motivation, affection, and support during these months. Thank you for all.

ABSTRACT

Mobility is one of the pillars of Smart Cities, being one of the most important issues for the development and growth of a city. Nowadays, with the massive flow of populations to large urban centers, mobility and its dangers require special attention. The number of pedestrian accidents has been increasing over the past few years, and it is important to understand what causes and factors contribute to them.

The main objective of this work is to identify and classify the pedestrian accidents in the city of Lisbon and segment the accident patterns based on the application of clustering methods. The data used in this work was provided by Lisboa Aberta.

The work begins with a literature review about the causes of pedestrian accidents previously identified and the reference of clustering methods used in similar studies.

Then, a deep dive will be made of the data provided by Lisboa Aberta, selecting the most relevant variables used for the Cluster analysis. The Cluster analysis will be done using the K-Means and K-Medoids methods.

At the end of the work, the results of both methods will be compared, where the winning method will be chosen and the conclusions of which are the determining patterns in pedestrian accidents in the city of Lisbon will be presented.

KEYWORDS

Pedestrian Accidents; Lisbon; K-Means; K-Medoids; Cluster

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ASNR	Autoridade Nacional Segurança Rodoviária
DGT	Direção Geral do Território
EU	European Union
ICT	Information and Communication Technologies
INE	Instituto Nacional de Estatística
LCC	Latent Class Clustering Analysis
RFE	Recursive Feature Elimination
SOM	Self-Organizing Maps
PAM	Partitioning Around Medoids
VIF	Variance Inflation Factor
WHO	World Health Organization

1. INTRODUCTION

Over the past centuries, cities have become places of great attraction for people and are considered as “magnets of hope’ for a vast array of skilled and unskilled people who flock to them to find better livelihoods and lifestyles” (Firoz and Kumar 2017). The movement of people from the countryside to the city has been a recurrent phenomenon, with a special emphasis in the last fifty years. The number of inhabitants in major cities around the world has been increasing over the past few years, reaching 55.3% in 2018 vs 34.0% in 1960, and the trend is to continue to grow, with an estimate of 60.4% in 2030 (Nations 2018).

The constant growth in terms of population and size in cities brings new challenges and the need to reinvent, evolve and put technology in their favor. Recently, and combined with new technologies and their integration, the concept of smart cities has emerged, which has several definitions: a) A smart city is a well-defined geographical area, in which high technologies such as ICT (information and communication technology), logistic, energy production, and so on, cooperate to create benefits for citizens in terms of well-being, inclusion and participation, environmental quality, intelligent development (Dameri 2014); b) All in all, smart city is the product of digital city combined with the Internet of Things (Su, Li, and Fu 2011); c) Smarter Cities are urban areas that exploit operational data (traffic congestion data, power consumption statistics, and public safety events), to optimize the operation of city services. The foundational concepts are instrumental, interconnected, and intelligent (Harrison et al. 2010).

The concept of smart cities can be divided into six parts: Smart Economy, Smart Mobility, Smart Governance, Smart Living, Smart People and Smart Environment (Giffinger et al. 2007). This study addresses one of the most complex and challenging pillars of a city, mobility, more specifically pedestrian mobility. The mobility in large cities is often characterized by car traffic, constant traffic issues, clutter and air pollution. To avoid this problem, governments and local authorities have been investing and promoting pedestrian mobility, which correctly combined with safety issues, contributes to more environmentally sustainable mobility, physical exercise of its inhabitants and can help mitigate local traffic.

Despite the environmental and health benefits of walking, the pedestrians sometimes are exposed to a higher risk of injury and fatality in road crashes. Pedestrians are often referred to as "vulnerable road users” since they are the most vulnerable victims in a road accident. Unprotected by vehicle body, safety belts or helmets, they are especially exposed to risk of serious injury and have a smaller chance of surviving an accident (Olszewski et al. 2015). In 2016, 5.320 pedestrians were killed in pedestrian accidents in the European Union (which is 21% of all road fatalities), while in Portugal the number was 123 deaths (European Commission 2018). According to World Health Organization, the annual fatality of traffic accidents worldwide reached 1.35 million each year, of which pedestrians and cyclists account for approximately 26% (WHO 2018).

The importance of pedestrian mobility contributed to the emergence of the walkability concept that can be characterized by “the extent to which the built environment is friendly to the presence of people walking, living, shopping, visiting, enjoying or spending time in an area” (Burton 2010), which demonstrates the concern of the pedestrian flow in the cities. However, pedestrian safety is an important social problem, which has not received the necessary attention. Nowadays, with the

available technologies and with the data produced and stored in a city, there are several variables that can be considered as relevant to explain the causes of an accident. It is important to understand which factors are correlated with accidents, to take a proactive attitude in the defense and safety of citizens. It is also important to identify the places where a greater number of accidents occur, usually called hotspots (Anderson 2009). The identification of a hotspot plays a fundamental role in the prevention of future accidents and should not only alert the danger of the local, but also promote safety changes.

The main objective of this study is to identify the factors that could be related with pedestrian accidents in the city of Lisbon on the following two points: 1) Analyze and classify the areas with the highest number of pedestrian accidents; 2) Investigate the correlation between pedestrian accidents and built environment factors including land use patterns, population, road infrastructure and transit characteristics through cluster analysis, namely K-Means and K-Medoids.

2. LITERATURE REVIEW

Accidents are unpredictable most of the times, being numerous times associated with human and mechanical causes. However, and with the development of technology, spatial factors are no longer underestimated and began to play an important role in the identification and prevention of pedestrian accidents (Whitelegg 1987). The comparison of accident spots, the identification of area characteristics and the segmentation of victims have been fundamental processes in the identification of patterns in pedestrian accidents.

The following literature review is divided into two parts: a first part with the most relevant factors identified in previous studies, and a second part with an explanation of the methodology used for the study, cluster analysis.

2.1. CAUSES OF PEDESTRIAN ACCIDENTS

In recent years, studies on mobility in urban areas have increased, and based on the most recent studies, it is possible to divide the causes related to pedestrian accidents into five categories: I) Land Cover; II) Socio-Economic characteristics; III) Roadway characteristics, IV) Victim characteristics and V) Weather conditions.

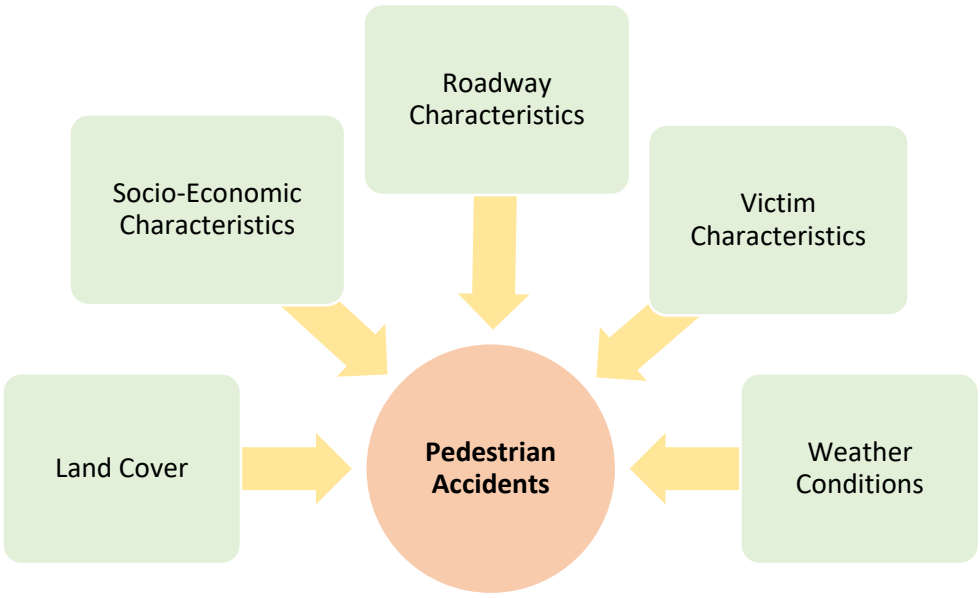


Figure 2.1 - Conceptual framework: Explanatory factors of pedestrian accidents

2.1.1. Land Cover

Before entering in a micro analysis on the exact location of the accident, it is important to understand in what type of area it happened, through the macro analysis of the location of the event. Pedestrian accidents occur in different areas, and sometimes the details of the location (characteristics of the road, speed limit, the presence of crosswalks, among others), overlap the area in question (Morency and Cloutier 2006). It is necessary to analyze an accident, from a macro scenario to a micro scenario.

Historically there are certain areas with a greater preponderance for the occurrence of accidents. Recent studies indicate that pedestrian accidents occur more frequently in urban areas (Hebert Martinez and Porter 2004). Areas with educational institutions are positively related to pedestrian accidents (Loukaitou-Sideris, Liggett, and Sung 2007; Yueying Wang et al. 2013). The students are usually children, and children are exposed to the dangers of traffic, as they tend to run and play along the roads and are often inexperienced with traffic rules. Their reaction time is also reduced, and they are unaware of the dangers. In addition to school zones, commercial areas (Kinga Ivan et al. 2015; Loukaitou-Sideris, Liggett, and Sung 2007; Tay and Rifaat 2007), industrial areas (Tay and Rifaat 2007), and residential areas stands out (D. J. Graham and Glaister 2003). Through the development of a negative binomial spatial model (D. J. Graham and Glaister 2003), found that fatalities were more likely in residential areas than in commercial areas. Other types of areas, such as retail and parks also have a positive correlation with pedestrian accidents (Kim, Made Brunner, and Yamashita 2006; Pulugurtha, Duddu, and Kotagiri 2013).

In terms of population density and car traffic, there is not just one theory: the first theory argues that the areas with the highest car volume are positively related to pedestrian accidents (Dumbaugh and Li 2011); on the other hand, a negative correlation between population density and the number of accidents was found (D. J. Graham and Glaister 2003). Pedestrian safety increases with the presence of more people, with speed limit and traffic control devices (Jacobsen 2015). The author also concludes that the presence of more people on the street reduces the risk of an accident, since drivers are more aware of the danger. Also, the speed limit increases the risk of an accident.

2.1.2. Socio-Economic characteristics

In addition to defining the typology of the accident area, socio-economic characteristics represent an important weight in identifying the area's most likely to have accidents with pedestrians. Several studies indicated a positive relationship between pedestrian accidents and low-income areas and a negative relationship with areas with higher income (Loukaitou-Sideris, Liggett, and Sung 2007; Tay et al. 2011). As a result, children with lower incomes are exposed to an increased risk of accidents (LaScala, Gruenewald, and Johnson 2004). The average household income has been associated with reduced traffic crashes (Dong et al. 2014; Zeng and Huang 2014). Also note that the higher the unemployment rate in a given area, the greater the likelihood of pedestrian accidents (Huang and Chin 2010).

Education and formation also represent a determining factor in defining the population most likely to suffer accidents. Children whose parents or guardians have a job with less educational qualifications are more likely to be involved in pedestrian accidents (D. Graham, Glaister, and Anderson 2005). People with a low level of education have three times a higher probability of death caused by pedestrian accidents than people with a higher level of education (Moudon et al. 2011).

People without their own car tend to travel on foot, and the proportion of households without vehicles shows a positive relationship with car accidents (J. Lee, Abdel-Aty, and Jiang 2014). In relation to the immigrant population, it was initially considered that new inhabitants were associated with a greater risk of accident, compared to residents of the country (Levine, Kim, and Nitz 1995), since the association between ethnicity and accident probability is made through a consideration of

low income by immigrants. More recent studies prove that immigrants, once out of their comfort zone, tend to follow the rules in a more restricted way, increasing their security levels (Reed and Sen 2005).

2.1.3. Roadway characteristics

Previous research has focused on the relationships between roadway characteristics and pedestrian safety. Entering a micro level, it is important to understand the characteristics of the accident site to classify and catalogue which factors become recurrent in a pedestrian accident. It is necessary to understand the interaction between pedestrian crashes and various characteristics of the urbanized environment to improve the security of the citizens. The roadway characteristics includes roads infrastructures, pedestrian infrastructures, buildings, and streetscape.

Starting by analyzing the accident site, it is important highlight that the total road length has been positively associated with car accidents (Hadayeghi, Shalaby, and Persaud 2010; Yueying Wang et al. 2013). However, the speed limit of the location may affect the importance of the road length. The roadway length was positively associated with crashes throughout the speed limit range of 40–105 kilometers per hour (Abdel-Aty et al. 2013). The speed limit in the area is strongly related to the likelihood of accidents (Sze and Wong 2007), as well as a higher risk of mortality and severity injuries. The average annual daily traffic for trucks is a statistically significant crash predictor (Huang and Chin 2010). In addition to the road length, the number of intersections is correlated with the number of accidents. The greater the number of intersections and the greater their distance, the greater the likelihood of accidents (Abdel-Aty et al. 2013; Dong et al. 2014; Huang and Chin 2010; J. Lee, Abdel-Aty, and Jiang 2014). A study performed with a logistic regression analysis, conclude that the probability of a crash is almost two times more likely at a site using traffic control than at a site without control (Ossenbruggen, Pendharkar, and Ivan 2001). The presence of crosswalks has been a depth topic in previous studies. Recent studies defend the idea that the crosswalks should not be painted without additional safety measures, since it transmits a false idea of safety to pedestrians and may not be transversal to drivers. Also, the probability of a crash is approximately two times more likely at a site without a sidewalk than at a site with a sidewalk (Damsere-Derry et al. 2010). The quality of pedestrian crossings design may significantly reduce the perception of being involved in an accident (Bernhoft and Carstensen 2008).

The built-up environment of the street directly affects pedestrians' perceptions of the spatial quality of the street and determines their activities (Zhang, Zhang, and Yin 2021). Recent studies try to understand the relationship between buildings and pedestrian mobility and safety. It was identified five major dimensions of the built environment which are determined for pedestrian mobility. They have known as the five D: "Density, diversity (land use mix), design (including street connectivity), distance to transit, and destination accessibility" (Cervero et al. 2009). In architectural terms, construction works and changes in buildings can impact pedestrian accidents (Richard A Retting, Susan A Ferguson, and McCartt 2003).

2.1.4. Victim characteristics

After understanding the factors surrounding an accident, it is necessary to understand the factors, both the driver and the victim. Although an accident can be unexpected, there are certain factors that seem to draw patterns. Age, gender, the presence of alcohol and drugs, the type of vehicle, are all factors to consider. Analyzing the age group, several studies agree that young people and children are more exposed to accidents than adults (Al-Ghamdi 2002; Johnson et al. 2004). In the case of children, the influence of parents on children's road education is a crucial factor (Pfeffer, Fagbemi, and Stennet 2010). Likewise, people over 60 years old are also at great risk of an accident (Abdel-Aty et al. 2013; Al-Ghamdi 2002; Ryb et al. 2007; Yueying Wang et al. 2013). This can be explained by factors such as lack of physical capacity, lack of precaution and lack of knowledge. Additionally, children and the elderly take more time to cross a road, increasing their exposure to danger (Demetriades et al. 2004). In terms of gender, (Sullman, Thomas, and Stephens 2012), men are more vulnerable in the circulation of a road, than women, otherwise some studies (Clifton et al. 2004), considers that women are more involved in accidents in areas with a higher population density.

The use of alcohol and drugs are two factors highly correlated with accidents, both in terms of the driver and the victim (C. Lee and Abdel-Aty 2005; Preusser et al. 2002; Ryb et al. 2007). The behavior of pedestrians and drivers is an important factor, since the lack of prudence and safety can lead to an accident (Miškinis and Valuntaite 2011; Ryb et al. 2007). In terms of vehicle type, passenger cars are more likely to be involved in pedestrian accidents (C. Lee and Abdel-Aty 2005).

2.1.5. Weather conditions

In addition to the four categories mentioned above, there is a topic that is gaining greater relevance in the definition of accidents with pedestrians: the weather conditions. Associated with the location, the meteorological conditions can play an important role in an accident. Weather conditions are one of the factors that contribute to pedestrian accidents and may influence the view of drivers and pedestrians (Kim, Pant, and Yamashita 2010). Heavy rain and sun (with air temperatures above 30 degrees) are correlated with pedestrian accidents (Li and Fernie 2010; Maze et al. 2008). Also, the increase and intensity of the rain are associated with loss of visibility on the part of the driver and pedestrians, increasing the risk of an accident (Theofilatos and Efthymiou 2012). More adverse temperatures make drivers and pedestrians less patient and more likely to violate traffic rules, increasing the risk of an accident (Naik et al. 2016). The light and the time of day affect the visibility of those involved, and drivers who drive in darker environments can result in difficulties in perceiving the danger of other drivers and pedestrians, as well as in visual performance and reaction time (Fylan et al. 2018). In broader and brighter environments, this risk decreases (Fountas et al. 2020).

Weather data often loses relevance for studies, as they are not collected or are collected incorrectly by the authorities and are often dependent on the judgment of the person who collects the data (Naik et al. 2016).

2.2. CLUSTER ANALYSIS

Cluster analysis has been a methodology used in recent studies on pedestrian accidents, since it allows to recognize accident patterns from the analysis of dataset of pedestrian accidents, and not to restrict the analysis to factors individually and arbitrarily chosen prior to the implementation of any method (Prato, Gitelman, and Bekhor 2012). In this type of problem, different types of clustering methods were used: Latent Class Clustering Analysis (LCC) to investigate the statistical relationship between pedestrian injury severity outcomes (Sun, Sun, and Shan 2019), LCC to identify the contributing factors and to explore the severity of bicycle accidents (Sivasankaran and Balasubramanian 2020); Self-Organizing Maps (SOM) to mapping patterns of pedestrian fatal accidents in Israel (Prato, Gitelman, and Bekhor 2012); Two-Step Cluster Analysis to investigate pedestrian accidents in Athens (Theofilatos and Efthymiou 2012); K-Means clustering to analysis of accident times for highway locations (Aljofey and Alwagih 2018), K-Means clustering algorithm to examine patterns of pedestrian involved crashes in Honolulu (Kim and Yamashita 2007), and K-Means to analyze road accident data (Kumar and Toshniwal 2015); Supervised association rules mining on pedestrian crashes in urban areas (Das et al. 2019).

3. DATA AND METHODS

3.1. STUDY AREA

This study will target the city of Lisbon. The capital of Portugal has an area of 100.05 km² and has the highest number of inhabitants, with a resident population of 507 220 in 2018 (Instituto Nacional de Estatística 2018). Lisbon is a city characterized by great population mobility, as it concentrates many services, commerce, leisure, and accommodation. In recent years, the population in the city of Lisbon and its surroundings has grown (2 827 050 in 2011 to 2 846 332 in 2018), contributing to greater pressure on systems and infrastructures. For that reason, it is noticeable that the governors are concerned with updating and improving mobility conditions in the city, with the current president of the Municipal Chamber of Lisbon, Dr. Fernando Medina, referring to the following: "Mobility is the goal of the next decade" (INE - Instituto Nacional de Estatística 2018). In addition to the number of inhabitants of the city, it is important to consider the number of tourists visiting Lisbon. In recent years, investment in infrastructure and facilities has been increasing. The number of accommodations for tourists, including hotels and local accommodations, increased from 511 units in 2015, to 916 in 2016, representing an increase around 80%. In addition, highlight the increase in the number of overnight stays in the city of Lisbon.

	# Total Tourist Accommodations	# Overnight stays in tourist accommodations per 100 inhabitants
2015	511	1 972
2016	563	2 192
2017	682	2 483
2018	765	2 602
2019	916	2 751

Table 3.1 - Capacity for tourist accommodation and overnight stays in tourist accommodations per 100 inhabitants in Lisbon, 2015-2019. Source: INE – Instituto Nacional de Estatística 2019.

Characterizing mobility in the city of Lisbon and considering the number of trips by main mean of transportation, the car comes in first place with 45.1% followed by walking with 29.8%. The average duration per trip is about 26 minutes, as well as the average distance travelled per trip is 9 kilometers (INE - Instituto Nacional de Estatística 2018). Between 2011 and 2018, 12 340 accidents with pedestrians were recorded in the district of Lisbon, resulting in 142 mortal victims (Autoridade Nacional Segurança Rodoviária 2018).

	2011	2012	2013	2014	2015	2016	2017	2018
# Pedestrian Accidents	1620	1439	1540	1463	1553	1546	1582	1597
# Mortal Victims	17	13	16	17	13	14	15	37

Table 3.2 - Number of pedestrian accidents and mortal victims in the district of Lisbon, 2011-2016. Source: ASNR – Autoridade Nacional Segurança Rodoviária, 2016).

3.2. DATA PREPARATION

The dataset used in this study contains the record of accidents with pedestrians that occurred in the city of Lisbon, between 2013 and 2018. During this period, the information of 420 accidents were collected and divided into 213 variables. The dataset provides information such as the location, through latitude and longitude, the date and time of the accident and weather conditions, namely temperature, precipitation, wind speed and humidity. In addition to the accident information, the dataset provides information about the socioeconomic characteristics of the area in which they occurred. These data allow to characterize different aspects about the population living in the area surrounding the accident. These data were provided by the Instituto Nacional de Estatística, during the Censos in 2011. The dataset also contains data about land cover, which allows to characterize the area, allowing to distinguish a commercial area, a residential area, or a school area, among others. These data were provided by the Direção Geral do Território (DGT), with reference to 2018. To complement the data above, more data about tourist establishments and local accommodation were shared, provided by Turismo de Portugal and data on points of interest in the city of Lisbon, provided by Open Street Maps. All the variables were allocated to the hexagonal grid to which they belong. All this work of integration and data sharing was carried out by Lisboa Aberta.

The four sets of data mentioned above were then added to a hexagonal grid, which delimits the city of Lisbon, where each hexagon had an area of 30 000 m^2. The reason that each hexagon has this area, is related to visualization effects and because it is the average area of the Censos blocks used for statistical purposes.

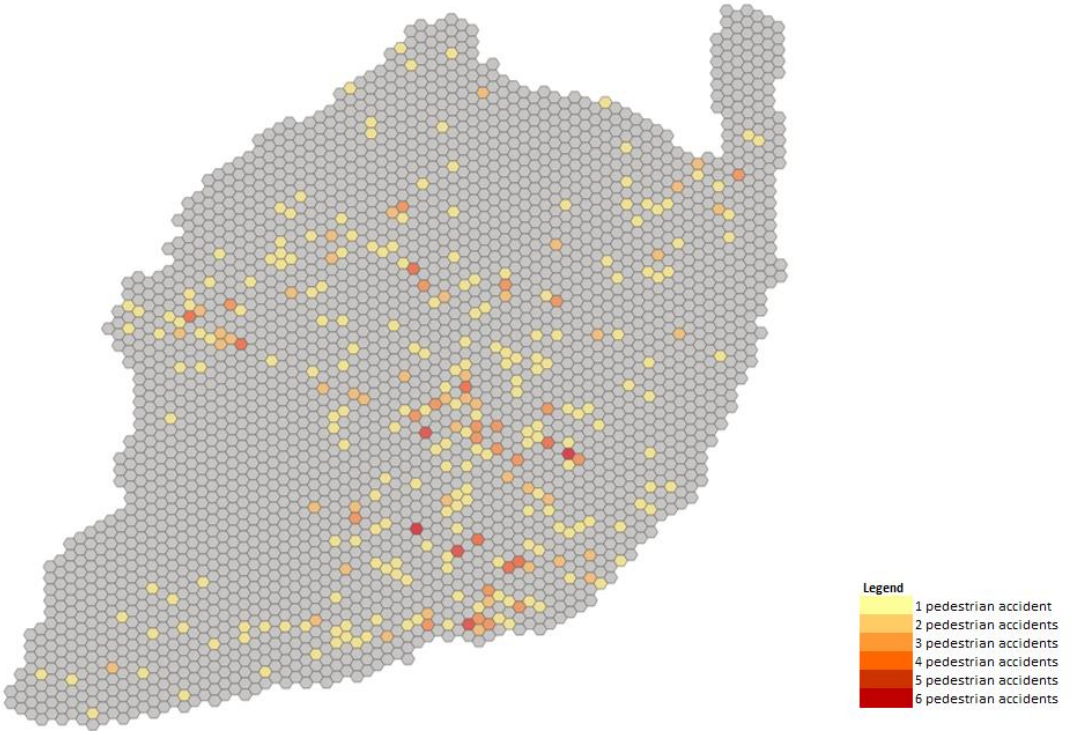


Figure 3.1 - Heat Map of the city of Lisbon with the registration of pedestrian accidents divided by the different hexagonal grids.

Performing an exploratory analysis of the data, it is useful to check the relevant variables, and identify some common factors and patterns in pedestrian accidents recorded in the dataset. In the next section, the variables will be grouped according to the groups identified in the literary review, where a summary of the most statistically relevant variables will be presented.

Note that the dataset does not include data about the characteristics of the victims and drivers, which makes it impossible for the study to consider the fourth category of the literary review: Victim characteristics. A renaming will also be made in the third chapter of the literature review, since the dataset does not include data about road networks and associated (e.g., crosswalks, traffic lights, number of roads, among others), having only information about the buildings around accidents area, and for that reason the name of the chapter will change to: Building Environment.

3.2.1. Pedestrian accident details

Analyzing the 420 accidents with pedestrians recorded in the city of Lisbon, it is possible to conclude through the dashboard of figure 3.2 the following conclusions: A) Graph A represents the distribution of the number of pedestrian accidents per year in the city of Lisbon. The year with the highest number of records was the year 2018, with 116 occurrences. From year to year, more pedestrian accidents were recorded; B) Graph B represents the distribution of the number of pedestrian accidents by season in the city of Lisbon. The season with the highest number of accidents is in autumn, with an occurrence rate of 33.1%, followed by summer with 24.5%. Winter is the season with the lowest number of cases, with a rate of 20.7%; C) Graph C represents the distribution of the number of pedestrian accidents by hourly intervals. The period of the day with the highest number of accidents registered is between 16 and 19 hours, in contrast, the period of the day with the least accidents is between 0 and 3 hours; D) Graph D represents the distribution of the pedestrian accidents per month. The months with the highest number of accidents are the months of October and September, with both having recorded 48 accidents. In contrast, February is the month with the fewest observations, however, it is also the month with the fewest days.

Based on the available data and drawing a profile, it is possible to conclude that accidents in the city of Lisbon are more likely to happen in the autumn, namely between September and November, with a special incidence in the afternoon, between 16 and 19 hours.

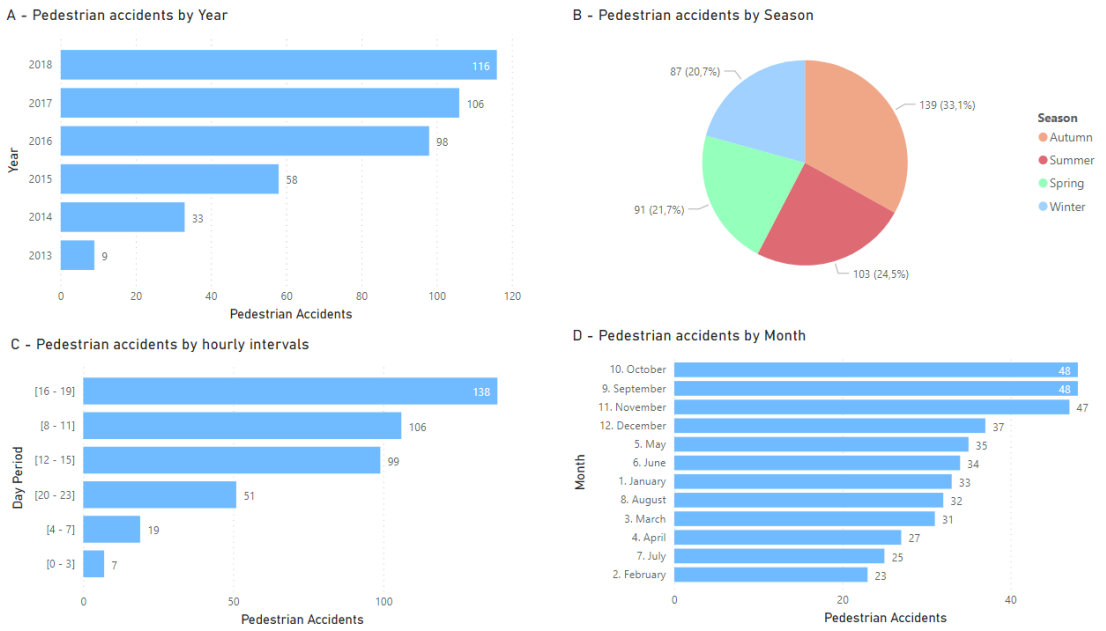


Figure 3.2 - Dashboard with relevant Pedestrian Accident Details

Considering the hexagonal grid, the 420 accidents recorded are divided into 287 distinct hexagons. This means that there are locations with more than one pedestrian accident. Table 3.3 shows the hexagons with the highest number of occurrences:

Grid ID	Longitude	Latitude	Zone	Nrº. of Pedestrian Accidents
BD-40	-9,127895443	38,73111977	Paiva Couceiro Square	6
AO-46	-9,155433355	38,72013210	Largo do Rato	6
AS-48	-9,149295493	38,71712395	Príncipe Real	5
AT-55	-9,146628092	38,70644769	Av. 24 de Julho	5
AP-38	-9,154889306	38,73394177	São Sebastião	5

Table 3.3 - Hexagon grid with the highest number of pedestrian accidents.

3.2.2. Land Cover

The dataset provides a set of 47 classification variables with the distribution in square meters of each service/category in the hexagon where the accident occurred. In this way it is possible to understand the composition of the 30 000 m^2 of each hexagon and proceed to its classification.

Considering the distribution of the land in the city of Lisbon presented on figure 3.3, the registered pedestrian accidents occurred in areas with predominantly vertical buildings with an average area of 15 430 m^2 (51.4%), which is natural, with Lisbon being a large urban center. The presence of road and associated networks, with an area of 5 320 m^2 (17.7%) is the second major factor to consider, having a relevant weight in the areas of accidents. Finally, highlight the areas of the city with tourist attractions, with an average area of 2 114 m^2 (7.0%), land with a discontinuous building with 1 045

m^2 (3.5%), and parks and gardens with 1 044 m^2 (3.5%). Variables with an average value of zero were not considered for the graph since they have no relevance in defining the areas of pedestrian accidents.

The combination of these numbers indicates that the areas of pedestrian accidents are mostly urban areas, with the presence of many buildings, related with an extensive road network. Associated with this, highlight the importance of tourist attractions and parks and gardens, in areas with a greater propensity for pedestrian accidents. This type of areas is characterized by high pedestrian mobility, which may be associated with recorded accidents.

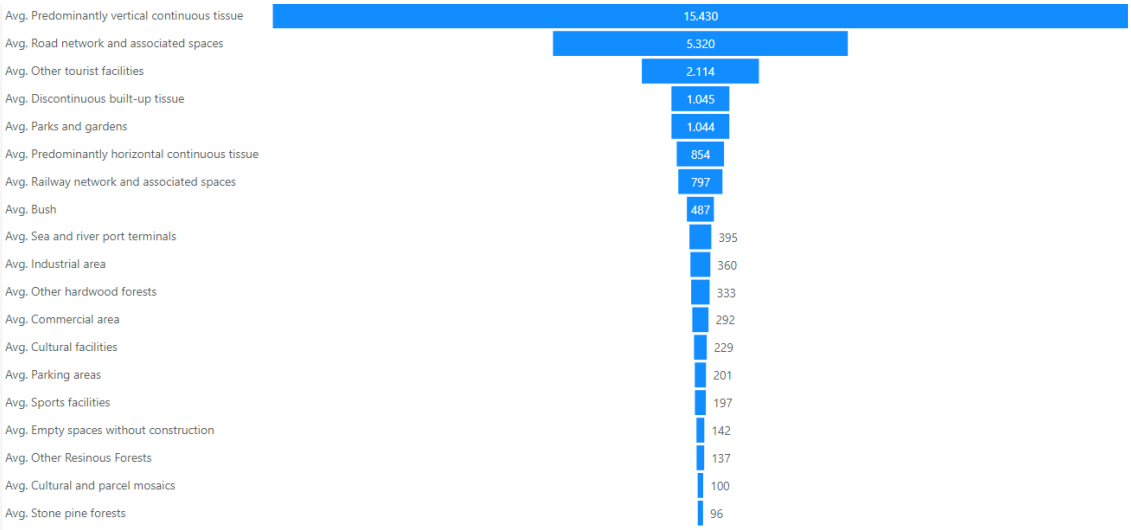


Figure 3.3 - Average distribution of the 30 000 m^2 of hexagons by Land Cover.

3.2.3. Socio-Economic characteristics

The socioeconomic variables available, allow to characterize the accident areas with the following highlights: A) Graph A represents the average distribution by gender of the hexagons where pedestrian accidents were recorded. Considering the gender of the resident population, emphasis is given to a greater ancestry of women (54.8% of resident women against 45.2% of resident men); B) Graph B represents the average distribution of the work activity of the hexagons where pedestrian accidents were recorded. On average, accident areas have a total of 294 inhabitants, of which 42.3% are employed inhabitants, 5.3% are unemployed and 4.2% are retired or pensioners. The rest of the population not characterized, is divided between population without economic activity and students; C) Graph C represents the average age distribution of the inhabitants of the hexagons where pedestrian accidents were recorded. Considering the age distribution, about 55.3% of the population is between 25 and 64 years old, with the second large group, corresponding to people over 65 years old, about 25.0%. The population considered young, under 25 years old, weighs about 19.7%; D) Graph D represents the level of education of the inhabitants of the hexagons where pedestrian accidents were recorded. Looking to the education of the resident population, 34.1% of residents completed university education, followed by 20.1% who only completed the 1st cycle. Highlight for a small fringe of the population (2.6%) of illiterate residents,

In general terms, the city of Lisbon can be characterized at an educational level with two extremes, since it has two representative groups, one with people with higher education and others who have just completed the first phase of education. In terms of age distributions, Lisbon residents are mostly adults or the elderly.

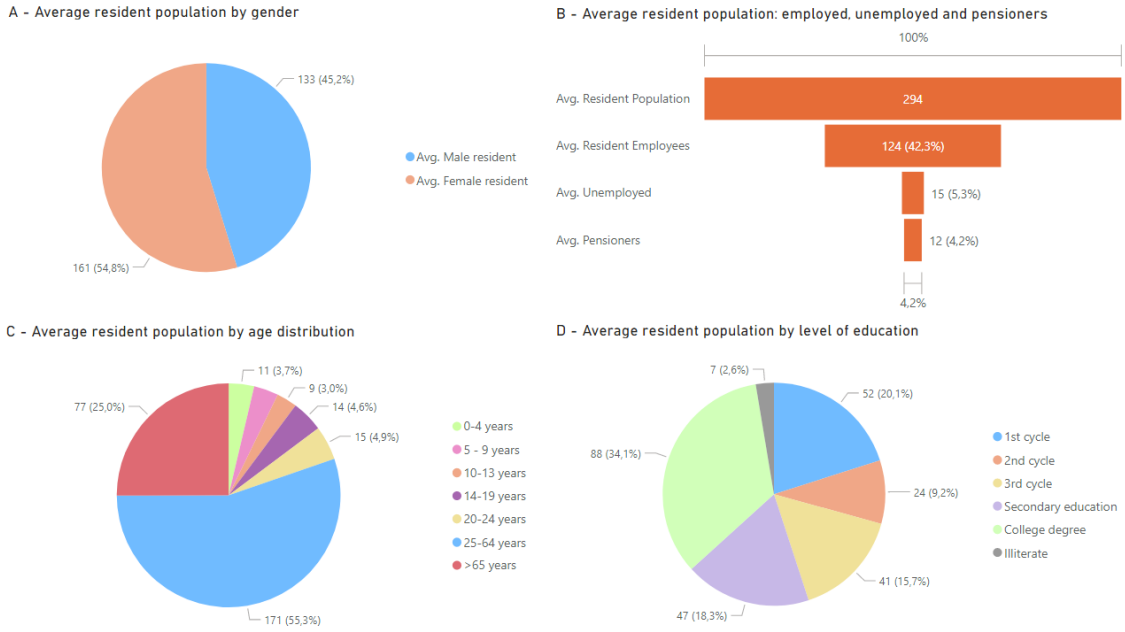


Figure 3.4 - Dashboard with the most relevant Socio-Economic variables

3.2.4. Building Environments

Considering the most relevant variables on the characteristics of buildings in accidents areas, it is possible to draw the following conclusions: A) Graph A represents the number and the type of households where pedestrian accidents were recorded. On average, 99.6% of buildings are classic family accommodation, such as apartments and houses. Only 0.4% correspond to collective accommodation, such as hotels and local accommodations. For houses without conditions, considered as tents, the percentage is less than 0.1%: B) Graph B represents the average number of floors of buildings in the hexagons where pedestrian accidents were recorded. Considering the height of the buildings 74.4% has 3 or more floors, with the most representative class being in 5 or more floors, 38.5%; C) Graph C represents the average number of buildings per year of construction. Analyzing the year of construction of the buildings in the areas surrounding pedestrian accidents, we conclude that about 66.1% of the buildings were built before 1960. The data set has records of the buildings built before 1919 until 2011.The most common period of construction was between 1919 and 1945, with about 24.0%. On the other hand, only 1.98% of the buildings were built between 2006 and 2011; D) Graph D represents the average number of different types of buildings in the hexagons where pedestrian accidents were recorded. Regarding the layout and distribution of the buildings, 56.1% follow a band construction, which means that the buildings are constructed in a sequential manner. Also noteworthy is the number of semi-detached buildings, about 19.7%.

In summary, the construction of areas surrounding pedestrian accidents may be characterized by buildings of classic and family accommodation, with a medium-high structure, typically old buildings and follow a band distribution.

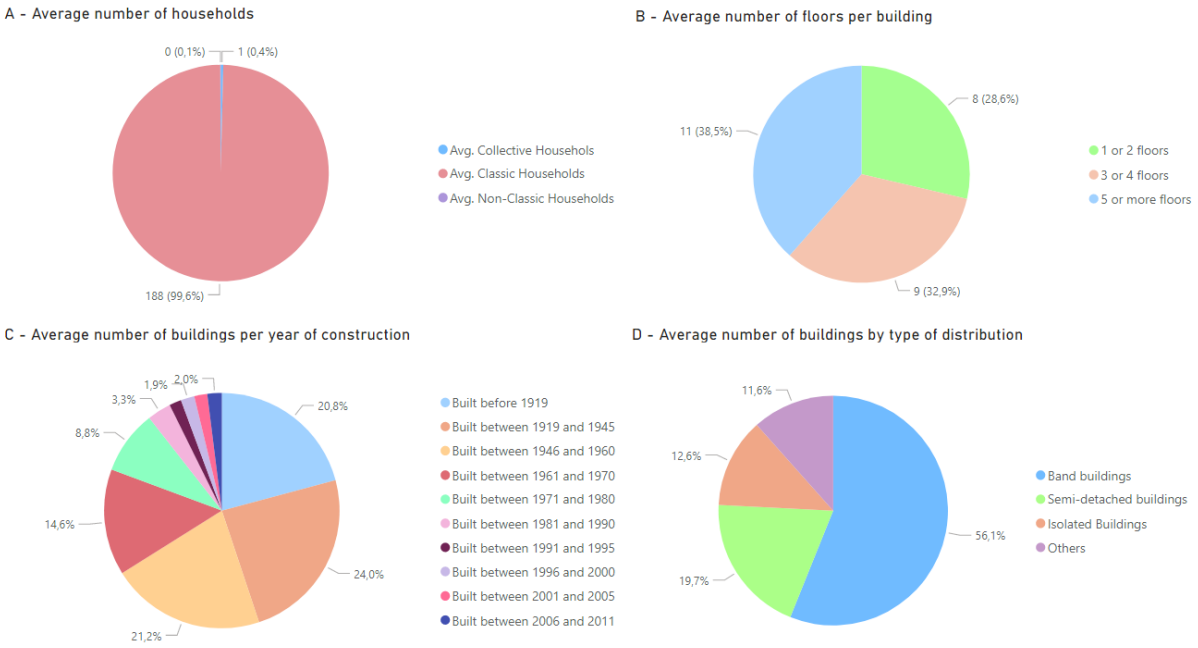


Figure 3.5 - Dashboard with the most relevant Building Environment's variables.

3.2.5. Weather conditions

Considering the variables available on the weather conditions on the day and in the pedestrian accident area, it is possible to draw the following conclusions: A) Graph A represents the distribution of the number of pedestrian accidents by temperature range, expressed in Celsius degrees. Most accidents were recorded with an average temperature ranging between 14º and 17º Celsius degrees. When considering the mild temperatures, varying between 11º and 17º Celsius degrees, 239 observations were recorded, which represents 56.9% of the total sample. Extreme temperatures register few observations; B) Graph B represents the average humidity recorded on the date of the accident, being expressed in a scale between 0 - 100%. It is possible to observe that pedestrian accidents are associated with periods with a medium/high humidity rate; C) Graph C represents the average wind speed recorded on the date of the accident and is expressed by a knot scale, and all values below 8 knots are considered light winds (Instituto Português do Mar e Atmosfera 2020). Only one occurrence of the dataset, registered moderate wind. The wind does not represent an explanatory factor in accidents with pedestrians in the city of Lisbon; D) Graph D represents the average rainfall recorded on the date of the accident and the precipitation is measured in millimeters. Through the data collected, it is possible to concluded that the accidents occurred on days with low or zero precipitation. Precipitation is not an explanatory factor.

Summing up the meteorological conditions, we conclude that the accidents with pedestrians in the city of Lisbon, occur on days with mild temperatures (between 11º and 17º degrees Celsius), with high humidity values, low wind speed and without precipitation.

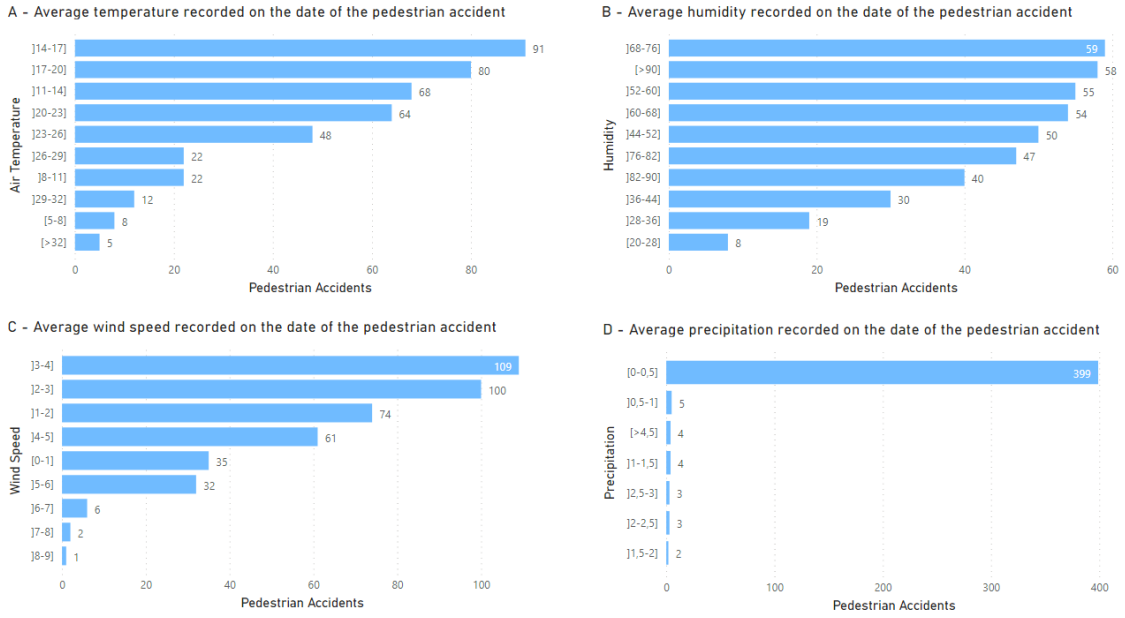


Figure 3.6 - Dashboard with Weather Conditions recorded at the date of accidents with pedestrians

4. METHODOLOGY

The dataset contains many variables, which can add some complexity to the study, so it is necessary to clean and remove the variables that are less relevant and that have the least influence on the study. Therefore, the dataset will be grouped by literature review group's and will go through three different stages: in a first phase, all variables that do not register any value will be removed; in a second phase, the Recursive Feature Elimination (RFE) technique will be applied; and finally, the third phase with the application of the Variance Inflation Factor (VIF) methodology.

After processing the data, the analysis chosen to extract results and subsequently knowledge, will be done through the application of a cluster analysis, namely through the following algorithms: K-Means and K-Medoids. All the methods mentioned above will be performed through the R Studio program, since it allows access to different packages and execute several models and algorithms in a single program. To estimate and identify the ideal number of clusters, different methods will be used, described during the chapter.

4.1. RECURSIVE FEATURE ELIMINATION (REF)

The technique used to remove complexity from dataset and understand which are the most relevant (independent variables) in the prediction of pedestrian accidents (dependent variable), will be through the application of a multiple linear regression with the Recursive Feature Elimination (RFE), considering all the available variables (excluding the variables without data).

RFE consists of an algorithm that allows to determine in a rigorous way the most important and relevant variables of a dataset in the prediction of the dependent variable in a predictive model. The REF algorithm works in three stages: I) In the first phase, build a model with all the variables in the dataset and calculate the importance of each one in the model. The importance calculations can be model based e.g., the random forest importance criterion naïve Bayes, bagged trees, linear regression, and others; II) Organizes the variables in order of importance and iterate through by building models of given subset sizes, that is, subgroups of most important predictors determined from step I). Ranking of the predictors is recalculated in each iteration; III) In the last stage, the model performances are compared across different subset sizes to arrive at the optimal number and list of final variables. The quality of the different models can be measured by different indicators, like Root Mean Square Error, Accuracy or Kappa (Kjell Johnson and Kuhn 2013).

"RFE works by searching for a subset of features by starting with all features in the training dataset and successfully removing features until the desired number remains" (Kjell Johnson and Kuhn 2013).

4.2. VARIANCE INFLATION FACTOR (VIF)

After choosing the most relevant variables in the definition of the problem, it is important to remove surplus variables. In dataset with a lot of data and variables it is very common to have multicollinearity problems, that is, when an explanatory variable is strongly related to a linear combination of the other independent variables. Multicollinearity does not violate the assumptions

of the model, but it does increase the variance of the regression coefficients. As the dataset is composed of many variables, the approach chosen to resolve the issue of multicollinearity is through the variance inflation factors (VIF). The VIF for the j^{th} independent variable is given by:

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R^2j is the R^2 from the regression of the j^{th} explanatory variable on the remaining explanatory variables (Steven D. Brown, Tauler, and Beata Walczak 2009). The VIF of an explanatory variable indicates the strength of the linear relationship between the variable and the remaining explanatory variables. A rough rule of thumb is that the VIFs greater than 10 cause for concern (Forthofer, Lee, and Mike Hernandez 2007). The definition of a model with fewer variables is important in this study to perform a Cluster analysis. For this study and follow the most recent literature, the VIF value to be considered will be 10.

4.3. CLUSTER ANALYSIS

Cluster analysis is a powerful statistical technique, which given a set of data, allows the identification of groups with similar behaviors and characteristics, and may reveal patterns related to the phenomenon under study (Julien Boccard and Serge Rudaz 2013). Unlike other tools, clusters groups observations instead of variables. In their composition, clusters must maximize the distance between clusters and minimize the distance within each cluster (Pang-Ning Tan et al. 2004). In the clustering process, there is a vast number of methods available. The most common are hierarchical methods, non- hierarchical methods, model bases methods and methods that allow overlapping clusters. Once the method is chosen, it is also possible to differentiate by the algorithm applied (D. Magnusson and Bergman 2001).

4.3.1. K-Means

For this study, a partition method was chosen, which will divide the dataset into *k*-clusters without any hierarchical relationship, K-Means' method. K-Means is one of the most common methods in cluster analysis, since: I) It is simple to implement and allows easy interpretability of the created clusters; II) Works well with large datasets; III) Generalizes to clusters of different shapes and sizes, such as elliptical clusters; IV) Guarantees convergence; V) Works well with continuous data.

The K-Means algorithm is an iterative process, which partitions the dataset into *k* clusters. Each point in the cluster is placed in the cluster closest to the cluster's mean value, which is called centroid (Sterling, Anderson, and Maciej Brodowicz 2018). The number of clusters (*k*) are specified by the user and $k \leq n$. The algorithm classifies the data into *k* groups, by satisfying the following requirements: each group contains at least one point, and each point belongs to exactly one cluster. The algorithm follows the following steps: I) Given *k*, the partition method creates an initial partition, typically randomly, and choose seeds; II) For each point in the dataset, calculate the distance between the point and all centroids and the point will be assigned to the cluster with the nearest centroid; III) Update the value of the centroid with the new mean value; IV) Repeat step II and III; V) End when the centroids cease to be recentered.

This process is repeated until all objects are classified within the multiple groups and the variation within the clusters is minimized by the sum of squared error (Hawas and Guo 2019). To calculate the average value of each cluster and to compute the distance of each point from the matched cluster, which is the nearest cluster, the algorithm used in this project will be the Euclidean distance:

$$W(C_k) = \sum_{\chi_i \in C_k} (\chi_i - \mu_k)^2$$

Here, x_i represents a point that belongs to the C_k cluster, and $U_k$represents the average of the value assigned to the C_k cluster. Each observation x_i , is assigned to a cluster so that the sum of the squares of the observation distance from its central cluster (U_k) is minimal. To validate the intra-cluster variation, the following formula is used:

$$tot.instracluster = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{\chi_i \in C_k} (\chi_i - \mu_k)^2$$

However, this method has three major limitations: I) it assumes prior knowledge of the data and requires the analyst to choose the appropriate number of cluster (*k*) in advance; II) The results obtained are sensitive to the initial random selection of cluster centers; III) It is sensitive to outliers. For the first limitation presented, it will be computed K-Means for a range of *k* values, and then, choose the best *k* by comparing the clustering results obtained for the different *k* values. For the second limitation, the K-Means algorithm will be tested several times with different initial cluster centers. The run with the lowest total within-cluster sum of square is selected as the final clustering solution. To overcome the problem of outliers, another algorithm, K-Medoids, will be tested, to be able to compare with the results of K-Means and understand if the problems with outliers will be solved.

4.3.2. K-Medoids

K-Medoids is an algorithm that is very close to K-Means, since both are partition algorithms and, in both cases, the objective is to group the different observations into *k* groups, with the *k* value being defined a priori. While in K-Means each cluster is represented by the average of the points present in the cluster and the objective is to minimize the sum of the square errors, in K-Medoids, clusters are represented by one of the points located near the to the center of the cluster (medoids), and the objective is to minimize the average dissimilarity of objects to their closest selected object (Park and Jun 2009).

To measure the dissimilarity between observations, the Gower Distance will be used for this study. Gower Distance proposal is the most popular way of measuring the similarity/dissimilarity between observations in the presence of mixed-type variable, allowing registers to be of different formats, such as numeric, categorical, logical or text (Marcello D'Orazio 2013). The distance is measured between 0 (identical) and 1 (maximum dissimilarity). The general formula is given by:

$$d_{G,j} = 1 - \mathcal{S}_{G,ij} = \frac{\sum_{t=1}^{\rho} \delta_{ijt}\, d_{ijt}}{\sum_{t=1}^{\rho} \delta_{ijt}}$$

The formula allows to calculate the distance or dissimilarity between observation i and observation j, where $d_{ijt} = 1 - S_{ijt}$ is the distance calculated on the t^{th} variable; $S_{ijt}$is the similarity between i and j with respect to the t^{th} variable and its value depends on the type of the variable.

After calculating the dissimilarity matrix, one of the most famous algorithms will be used to clustering, Partitioning Around Medoids (PAM). PAM algorithm searches for *k* representative objects in a data set (K-Medoids), and the operation of the algorithm is very similar to K-Means. In the step where the centroids are updated (step III), the K-Means were computing mean of all points present in the cluster. However, in the PAM algorithm, the process is as follows: for each medoid (C_i), and for each non-medoid data point (P_i), are considered the swap of C and P, and compute the cost change. Perform the best swap of C and P if it decreases the cost function. Otherwise, the algorithm ends. The cost in K-Medoids algorithm is given as:

$$c = \sum\sum|P_i - C_i|$$

When medoids are not specified, the algorithm first looks for a good initial set of medoids (this is called the build phase). Then it finds a local minimum for the objective function, that is, a solution such that there is no single switch of an observation with a medoid that will decrease the objective (this is called the swap phase)(Belhadi et al. 2020).

4.3.3. Elbow Graphic

After applying the cluster algorithm there are techniques that help to understand the composition of different clusters, the degree of quality of the different partitions and subsequently identify the ideal number of clusters for the study in question.

The first method to identify the optimal number of clusters is the Elbow Graphic. To define clusters where the within-cluster variation is as small as possible, this graph allows us to understand the variance within the clusters as the *k* value is adjusted. The formula applied is as follows:

$$minimize\ (\sum_{k=1}^{k} W(C_k))$$

where C_k is the k^{th} cluster and $W(C_k)$ is the within-cluster variation. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

4.3.4. Silhouette width

The second mechanism to consider is the Silhouette width. This is a widely used measure for assessing the fit of individual objects in the classification, as well as the quality of clusters and the entire classification. A high average silhouette width indicates a good clustering (its value is comprised between 1 and -1 with a value of 1 indicating a very good cluster). The average silhouette method computes the average silhouette of observations for different values of *k* (Batool and Christian Hennig 2021).

4.3.5. Gap Statistic Method

The application of the Gap Statistic Method is possible in any cluster method. This method compares the total within cluster variation for different values of *k* with their expected values under null reference distribution of the data. For each variable, an interval between the minimum and maximum is calculated, and generate values for the *n* points uniformly from the interval minimum to maximum (Tibshirani, Walther, and Hastie 2001).

4.3.6. NbCLust package

The fourth method to be used is based on a package of functions provided by the R Studio software, called NbCLust package. This package allows you to perform cluster analysis with 26 different indicators, helping to determine the optimal number of clusters.

The present methods are as follows:

Method	**Article**	**Method**	**Article**	**Method**	**Article**
KL	Krzanowski and Lai 1988	Duda	Duda and Hart 1973	Hubert	Hubert and Arabie 1985
CH	Calinski and Harabasz 1974	PseudoT2	Duda and Hart 1973	SDindex	Halkidi et al. 2000
Hartigan	Hartigan 1975	Beale	Beale 1969	Dindex	Lebart et al. 2000
CCC	Sarle 1983	Ratkowsky	Ratkowsky and Lance 1978	SDbw	Halkidi and Vazirgiannis 2001
Scott	Scott and Symons 1971	Ball	Ball and Hall 1965	Rubin	Friedman and Rubin 1967
Marriot	Marriot 1971	PtBiserial	Milligan 1980, 1981	Cindex	Hubert and Levin 1976
TrCovW	Milligan and Cooper 1985	Frey	Frey and Van Groenewoud 1972	DB	Davies and Bouldin 1979
TraceW	Milligan and Cooper 1985	McClain	McClain and Rao 1975	Silhouette	Rousseeuw 1987
Friedman	Friedman and Rubin 1967	Dunn	Dunn 1974		

Figure 4.1 - NBClust Package Methods

5. RESULTS AND DISCUSSION

5.1. VARIABLE SELECTION

The variable selection process will be carried out by grouping the variables by literary review group (variables with the characteristics of the victims not available). After the exploratory analysis of the entire dataset, the first step is to remove the variables without data or that do not fit the study; the second step is to apply the REF methodology to choose the most important variables. To apply the REF, linear models will be created in which the dependent variables of the literary review group aim to explain the dependent variable, pedestrian accidents in the city of Lisbon; the last step is to apply the VIF to remove the multicollinearity of the winning models.

Since the dataset data are in different scales, the application of the techniques referred to above, will be through the normalization of the data.

5.1.1. Land Cover

To describe the type of land in the city of Lisbon with a greater propensity for pedestrian accidents, the dataset provides 47 variables. After performing an exploratory analysis, 20 variables with agricultural and forestry data were excluded, since they are not considered relevant for the study of pedestrian accidents in large urban areas. Subsequently, variables with an average sample value of less than 500m^2 were also excluded, as they weigh very little in the definition of hexagons. With this exclusion, 20 variables are also excluded. The variables considered to the study are presented in Table 5.1.

Building the model with the selected variables and applying the REF, the variables chosen as the best to explain pedestrian accidents are as follows: Avg. Railway Network and Associated Spaces and Avg. Predominantly Vertical Continuous Tissue. The model with these two variables presents an R^2 of 74.0% The R^2, known as coefficient of determination, is a measure of adjustment of a generalized linear statistical model, such as simple or multiple linear regression, to the observed values of a random variable. The values vary between 0 and 1, and the closer the value is to 1, the greater the amount of variance in the data that is explained by the linear model (Brown, Lo, and Lys 1999).

After applying the VIF, the two selected variables do not reveal multicollinearity, since the value of both is less than 10.

Category	Variables	Mean	Median	Min.	Max.	Std.	VIF
Land Cover	Avg. Predominantly Vertical Continuous Tissue	15 430	17 046	0	30 000	10 516	1,2
	Avg. Predominantly Horizontal Continuous Tissue	854	0	0	30 000	3 794	-
	Avg. Discontinuous Built-up Tissue	1 045	0	0	30 000	4 179	-
	Avg. Road Network and Associated Spaces	5 320	4 791	0	30 000	5 317	-
	Avg. Railway Network and Associated Spaces	797	0	0	21 738	3 143	1,2
	Avg. Other Tourist Facilities	2 114	0	0	30 000	4 813	-
	Avg. Parks and Gardens	1 044	0	0	23 276	3 225	-

Table 5.1 - Variables to be considered in the model for the "Land Cover" group and VIF methodology.

5.1.2. Socio-Economic characteristics

The dataset contains 82 socioeconomic variables on the areas where pedestrian accidents were recorded. After an exploratory analysis of the data, 52 variables were excluded, since they were variables with no significance for the study (such as the structure of buildings in the city of Lisbon, the number of bathrooms in a building, among others), and variables with a high degree of granularity (for example, division of the complete level of education of men and women, division of the level of education to be attended among others and women, among others). At the end, 30 variables were chosen to be considered for the model, shown in Table 5.2.

Building the model with the selected variables and applying the REF, the variables chosen as the best to explain pedestrian accidents are as follows: Avg. Number of classic families with 2 or more unemployed, Avg. Number of individuals employed - primary sector, Avg. Number of individuals employed - tertiary sector, Avg. Number of individuals with complete education - Post-Secondary, Avg. Number of individuals with complete education - College degree, Avg. Number of individuals by age group: 10-13 years, Avg. Number of individuals by age group: 14-19 years and Avg. Number of classic family houses with parking for 3 vehicles or more. The model with these eight variables presents an R^2 of 83.1%

Category	Variables	Mean	Median	Min.	Max.	Std.	VIF
Socio-Economic	Avg. Number of classic families with 2 or more unemployed	1,4	0,8	0,0	19,8	2,0	2,9
	Avg. Number of classic families with 1 unemployed	12,2	8,6	0,0	67,4	11,3	-
	Avg. Number of classic families	137,9	112,7	0,0	469,1	117,7	-
	Avg. Number of institutional families	0,2	0,0	0,0	3,7	0,4	-
	Avg. Number of unemployed individuals	12,5	8,4	0,0	77,7	12,2	-
	Avg. Number of individuals employed - primary sector	0,4	0,1	0,0	3,2	0,5	1,3
	Avg. Number of individuals employed - secondary sector	12,4	9,9	0,0	49,7	11,5	-
	Avg. Number of individuals employed - tertiary sector	111,5	92,9	0,0	408,6	91,6	19,6
	Avg. Number of individuals employed	124,2	105,0	0,0	459,5	102,7	-
	Avg. Number of pensioners and retirees	80,8	56,8	0,0	404,7	77,2	-
	Avg. Number of individuals without economic activity	120,7	98,8	0,0	497,5	105,4	-
	Avg. Number of individuals with complete education - 1st cycle	52,0	33,9	0,0	283,6	53,2	-
	Avg. Number of individuals with complete education - 2nd cycle	23,8	17,9	0,0	142,7	21,8	-
	Avg. Number of individuals with complete education - 3rd cycle	40,6	31,1	0,0	156,9	37,2	-
	Avg. Number of individuals with complete education - post-Secondary	3,1	2,3	0,0	15,6	3,2	5,1
	Avg. Number of individuals with complete education - Secondary	47,5	35,8	0,0	205,4	42,3	-
	Avg. Number of individuals with complete education - College degree	88,3	70,9	0,0	340,7	80,9	7,0
	Avg. Number of illiterate individuals	6,8	3,4	0,0	71,8	8,9	-
	Avg. Number of individuals by age group: 0-4 years	11,4	9,3	0,0	75,1	10,4	-
	Avg. Number of individuals by age group: 5-9 years	11,0	9,1	0,0	80,7	9,9	-
	Avg. Number of individuals by age group: 10-13 years	9,2	8,4	0,0	53,7	8,1	9,1
	Avg. Number of individuals by age group: 14-19 years	14,1	13,1	0,0	75,7	12,0	10,2
	Avg. Number of individuals by age group: 15-19 years	12,1	11,0	0,0	65,9	10,3	-
	Avg. Number of individuals by age group: 20-24 years	15,3	13,1	0,0	62,7	13,1	-
	Avg. Number of individuals by age group: 25-64 years	155,6	132,7	0,0	554,1	128,1	-
	Avg. Number of individuals by age group: > 65 years	77,4	53,0	0,0	387,2	73,3	-

	Avg. Number of male individuals	133,0	115,3	0,0	496,7	108,2	-
	Avg. Number of female individuals	161,0	136,2	0,0	523,0	134,4	-
	Avg. Number of resident families	81,4	67,4	0,0	284,8	69,5	-
	Avg. Number of classic family houses with parking for 3 vehicles or more	1,9	0,7	0,0	120,6	7,0	1,4

Table 5.2 - Variables to be considered in the model for the "Socio-economic" group and VIF methodology.

After applying the VIF, there is multicollinearity in two of the selected variables: Avg. Number of individuals employed - tertiary sector and Avg. Number of individuals by age group: 14-19 years. After removing the two variables from the model, the R^2 goes from 83.1% to 84.2% showing an improvement for the remaining 6 variables.

5.1.3. Building Environments

The data about the characteristics of the buildings and associated services/attractions, brings together a total of 50 variables in the dataset. After an exploratory analysis of the data, 25 variables were excluded from the study, showing has no relevance to the analysis, and in the end 25 were considered for the model, shown in Table 5.3.

Building the model with the selected variables and applying the REF, the variables chosen as the best to explain pedestrian accidents are as follows: Avg. Number of households, Avg. Number of households with 1 or 2 floors, Avg. Number of households with 3 or 4 floors, Avg. Number of households with 5 or more floors, Avg. Number of buildings in band, Avg. Number of buildings semi-detached, Avg. Number of isolated buildings, Avg. Number of buildings: other, Avg. Number of buildings built before 1919, Avg. Number of bus stops, Avg. Number of metro stations, Avg. Number of restaurants, coffees, bars and supermarkets, Avg. Number of bike parks and Avg. Number of touristic establishments. The model with these fourteen variables presents an R^2 of 86.3%.

Category	Variables	Mean	Median	Min.	Max.	Std.	VIF
Building Environments	Avg. Number of households	188,4	144,5	0,0	653,8	156,7	7,3
	Avg. Number of households with 1 or 2 floors	8,0	1,4	0,0	125,5	17,5	38,4
	Avg. Number of households with 3 or 4 floors	9,1	3,0	0,0	74,8	12,6	3,0
	Avg. Number of households with 5 or more floors	10,7	7,7	0,0	52,4	10,3	5,0
	Avg. Number of buildings in band	4,3	0,2	0,0	105,2	13,1	21,1
	Avg. Number of buildings semi-detached	1,5	0,0	0,0	92,7	6,8	4,6
	Avg. Number of isolated buildings	1,0	0,1	0,0	53,8	3,3	2,8
	Avg. Number of buildings: other	0,9	0,1	0,0	15,6	2,4	2,7
	Avg. Number of buildings built before 1919	5,7	0,3	0,0	101,4	13,7	2,2
	Avg. Number of hospitals	0,0	0,0	0,0	1,0	0,1	-
	Avg. Number of health centers	0,1	0,0	0,0	2,0	0,3	-
	Avg. Number of pre schools	0,1	0,0	0,0	2,0	0,3	-
	Avg. Number of 1st cycle schools	0,1	0,0	0,0	2,0	0,3	-
	Avg. Number of 2nd and 3rd cycle schools	0,1	0,0	0,0	1,0	0,2	-
	Avg. Number of secondary schools	0,1	0,0	0,0	1,0	0,2	-
	Avg. Number of professional schools	0,0	0,0	0,0	1,0	0,1	-
	Avg. Number of universities	0,0	0,0	0,0	1,0	0,1	-

	Avg. Number of bus stops	1,6	1,0	0,0	6,0	1,4	1,1
	Avg. Number of metro stations	0,1	0,0	0,0	1,0	0,3	1,2
	Avg. Number of train stations	0,0	0,0	0,0	1,0	0,1	-
	Avg. Number of piers	0,0	0,0	0,0	1,0	0,1	-
	Avg. Number of cultural spots	0,1	0,0	0,0	2,0	0,4	-
	Avg. Number of restaurants, coffees, bars and supermarkets	2,7	1,0	0,0	29,0	4,6	1,4
	Avg. Number of bike parks	0,9	0,0	0,0	7,0	1,5	1,2
	Avg. Number of touristic establishments	0,2	0,0	0,0	5,0	0,7	1,2

Table 5.3 - Variables to be considered in the model for the " Building Environments " group and VIF methodology.

After applying the VIF, there is multicollinearity in two of the selected variables: Avg. Number of households with 1 or 2 floors and Avg. Number of buildings in band. After removing the two variables from the model, the R^2 goes from 86.3% to 87.3% showing an improvement for the remaining 12 variables.

5.1.4. Weather conditions and date

The last model considered for the analysis, includes the variables with the meteorological conditions, as well as the date and period of the day variables (these last two variables are categorical). The 6 variables are represented in Tables 5.4, Table 5.5, and Table 5.6.

Category	Variables	Mean	Median	Min.	Max.	Std.	VIF
Weather conditions	Temperature	18,5	18,1	5,3	35,9	5,6	7,3
	Humidity	66,8	68,0	21,0	100,0	19,1	38,4
	Precipitation	0,1	0,0	0,0	6,6	0,7	-
	Wind Speed	1 085,7	784,0	0,0	3 718,4	1 102,9	21,1

Table 5.4 - Variables to be considered in the model for the " Weather Conditions " group and VIF methodology

Variables	Fields	# Observations	Frequency (%)	VIF
Month	January	33	7,9%	4,6
	February	23	5,5%	
	March	31	7,4%	
	April	27	6,4%	
	May	35	8,3%	
	June	34	8,1%	
	July	25	6,0%	
	August	32	7,6%	
	September	48	11,4%	
	October	48	11,4%	
	November	47	11,2%	
	December	37	8,8%	
	Total	420	100,0%	

Table 5.5 - Variable “Month” to be considered in the model and VIF methodology

Variables	Fields	# Observations	Frequency (%)	VIF
Day Period	[0 - 3]	51	12,1%	2,8
	[4 - 7]	138	32,9%	
	[8 - 11]	106	25,2%	
	[12 - 15]	7	1,7%	
	[16 - 19]	99	23,6%	
	[20 - 23]	19	4,5%	
	Total	420	100,0%	

Table 5.6 - Variable "Day Period" to be considered in the model and VIF methodology

Building the model with the selected variables and applying the REF, the variables chosen were all except: Precipitation. The model with these five variables presents an R^2 of 83.6%

After applying the VIF, there is multicollinearity in two of the selected variables: Humidity and Wind Speed. After removing the variables from the model, the R^2 goes from 83.6% to 83.7% showing an improvement for the remaining variables.

After carrying out the exploratory analysis of the data, 24 variables were selected through the different groups of the literary review, as the most suitable to explain pedestrian accidents in the city of Lisbon. In the next step, the variables will be the target of the clustering process through K-Means and K-Medoids.

Of these 24 variables, 22 will be used in K-Means and 24 in K-Medoids. The number of variables selected is different since the limitation of K-Means in working with categorical variables (Day Period and Month). In the end, three models will be compared, K-Means with 22 variables, K-Medoids with 24 variables and K-Medoids with 22 variables to understand how it works without categorical variables. The two processes will be compared to identify which are the most relevant patterns to explain pedestrian accidents in the city of Lisbon, as well as which is the best cluster technique to apply to this study.

5.2. CLUSTER ANALYSIS

The next section will be divided into three scenarios: K-Means, K-Medoids with categorical variables and K-Medoids without categorical variables.

5.2.1. K-Means

K-Means is the first algorithm to be used in this study.

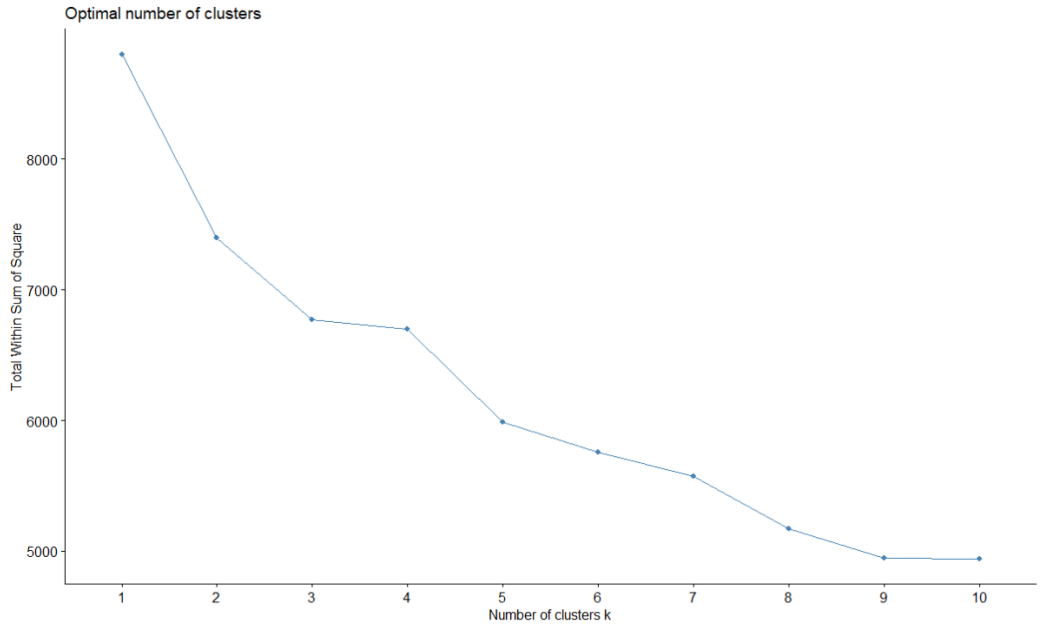


Figure 5.1 - Elbow Graph for K-Means

K	Total within-cluster sum of square	Variation (WCSSK - WCSSK-1)
2	16,0%	-
3	21,4%	5,4%
4	28,6%	7,2%
5	32,3%	3,7%
6	34,6%	2,3%
7	36,6%	2,0%
8	38,5%	1,9%

Table 5.7 - K-Means Evaluation Process

Analyzing the elbow graphic, the ideal number of clusters is not easily visible, since the ideal number can be *k*=2 or *k*=4. For that reason, it is necessary to consider more information, and for the K-Means algorithm the total within-cluster sum of square will be evaluated. Through the Table 5.7 it is possible to notice that as there are more partitions, the total within-cluster sum of square will increase. However, the biggest increase occurs when the *k* value goes from 1 to 2, with an increase of 16.0%.

The second mechanism to consider is the Silhouette width.

K	Average silhouette width
2	0,39
3	0,08
4	0,03
5	0,03
6	0,07
7	-0,06
8	-0,09

Table 5.8 - Average Silhouette for K-Means

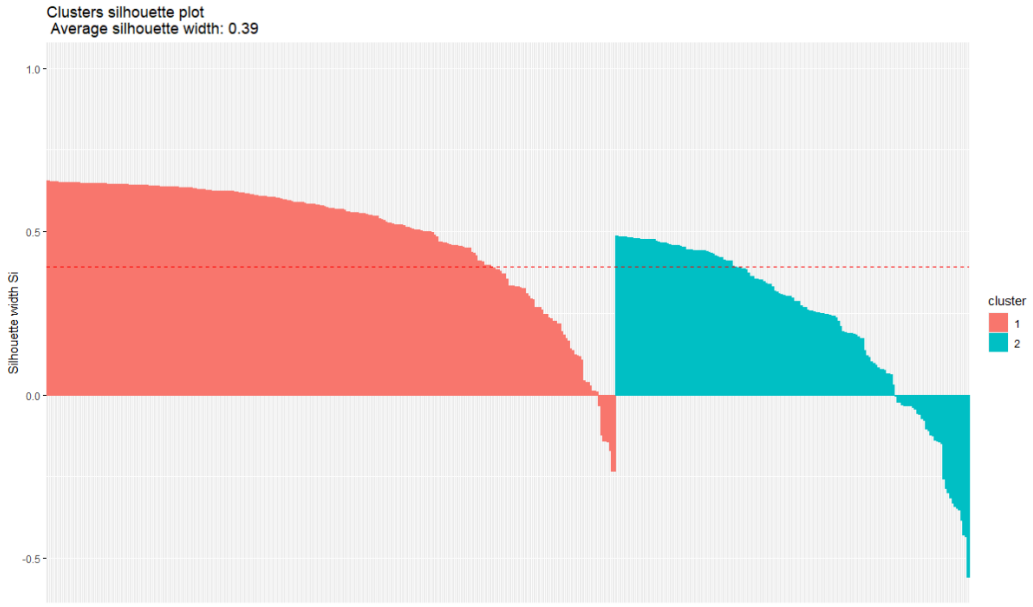


Figure 5.2 - Average Silhouette for *k*=2

Considering the second method, the k with the best classification is *k* =2, since it is the one with the highest value of average silhouette, revealing a better classification of objects when the dataset is split in two.

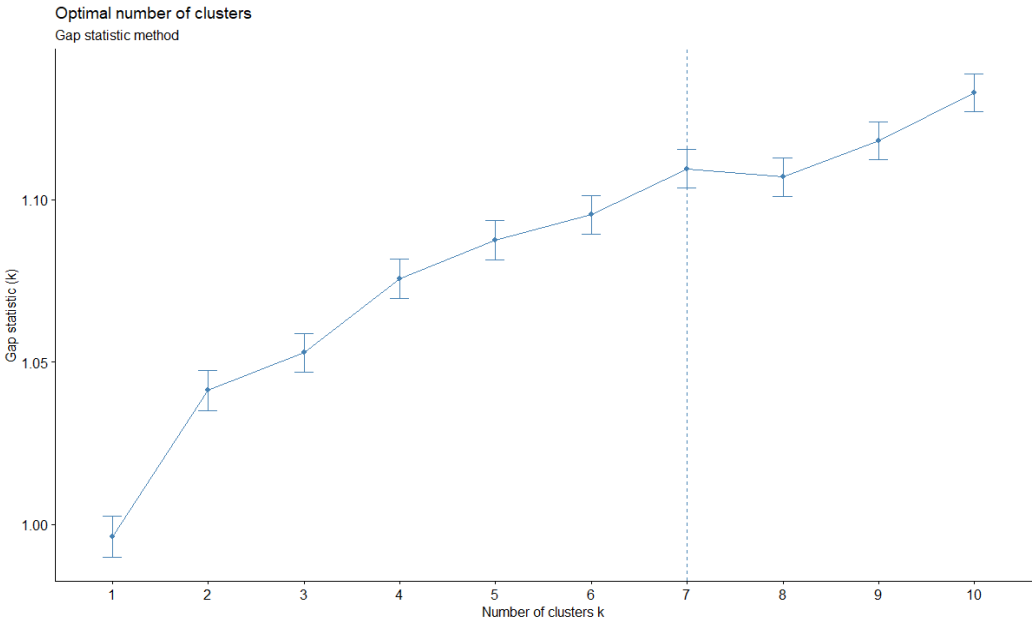


Figure 5.3 - Gap Statistic Method for K-Means

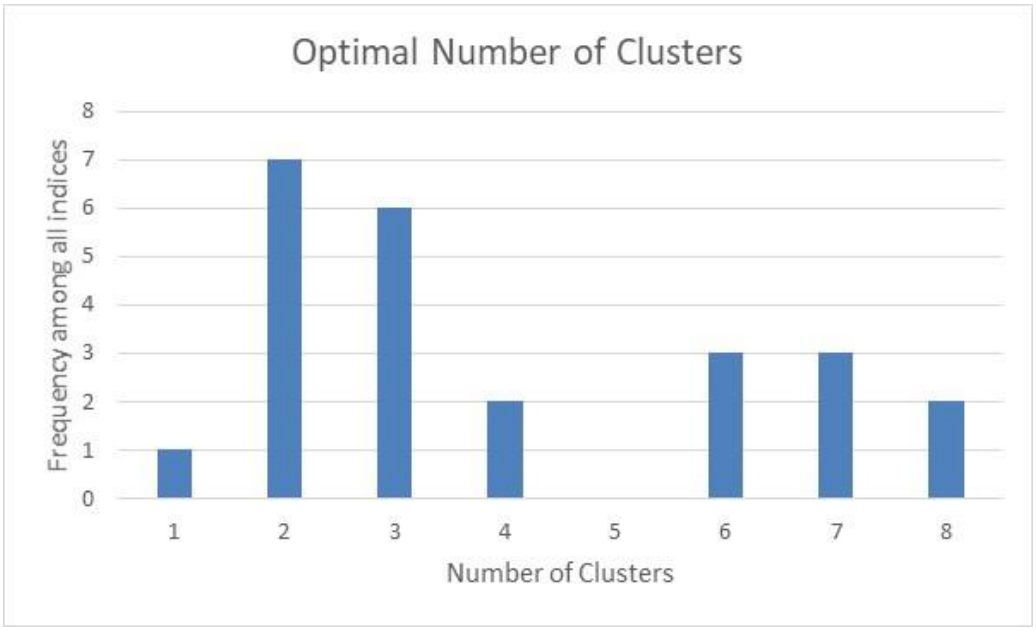


Figure 5.4 - NbCLust package for K-Means

Figure 5.3 represents the application of the third mechanism, Gap Statistic, which returns the following value for *k*=7. Using the fourth method, 26 different indicators were applied at once, with the most common partition value being *k*=2, being suggested by 7 times, Figure 5.4. Considering the different outputs, the value of *k* to consider for the K-Means is *k*=2, since corresponds to the partition where the within-cluster sum of square increase substantially, with the average silhouette higher and is the most common partition applying the NbCLust package.

Assuming the value of *k*=2, the first cluster will be composed of 259 observations, which corresponds to 61.6% of the sample, while the second cluster will be composed of 161 observations, which corresponds to 38.4% of the sample.

K /Cluster Dimension	1	2	3	4	5	6	7	8
2	259	161						
3	227	153	40					
4	189	34	75	122				
5	120	3	70	33	194			
6	136	11	40	30	140	63		
7	96	10	8	30	146	60	70	
8	95	10	8	18	129	53	77	30

Table 5.9 - Absolute Frequency of Clusters with K-Means Algorithm

5.2.2. K-Medoids without categorical variables

The second algorithm to be used for this study is the K-Medoids without the inclusion of categorical variables.

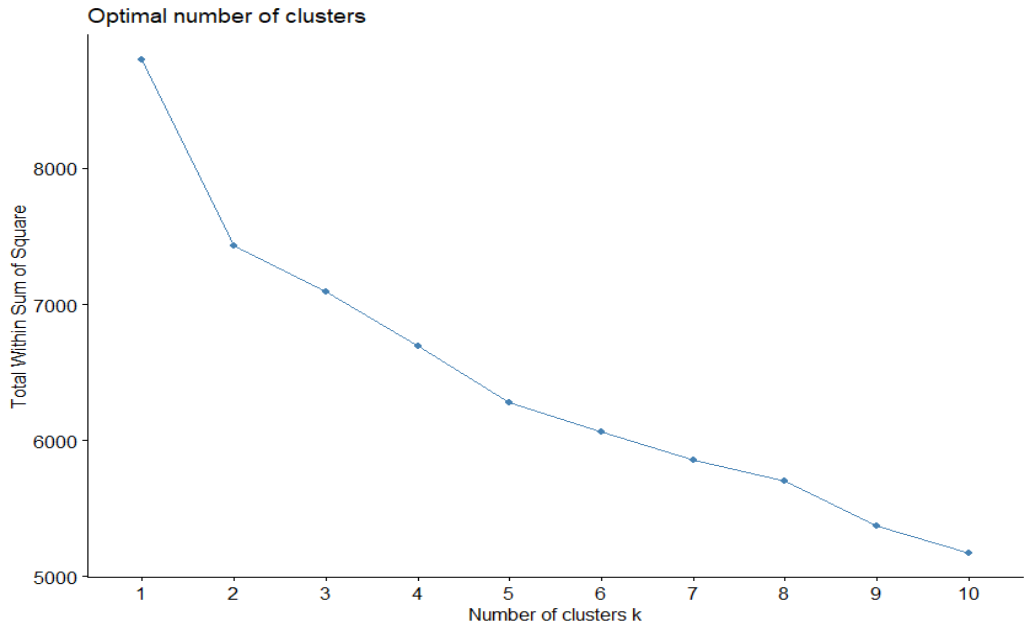


Figure 5.5 - Elbow Graph for K-Medoids

K	Build and Swap	Variation (B&SK - B&SK-1)
2	4,01	-
3	3,85	-16,3%
4	3,75	-10,0%
5	3,65	-10,0%
6	3,56	-9,0%
7	3,48	-8,0%
8	3,40	-8,0%

Table 5.10 - K-Medoids Evaluation Process

Considering the Elbow Graph in Figure 5.5 and K-Medoids evaluation process in Table 5.10 (the values listed are the values of the objective function, sum of distances of points to their medoid at the two stages and they represent a measure of how well the points clustered), the greatest variation occurs when *k*=2.

K	Average silhouette width
2	0,51
3	0,42
4	0,44
5	0,42
6	0,37
7	0,27
8	0,23

Table 5.11 - Average Silhouette for K-Medoids

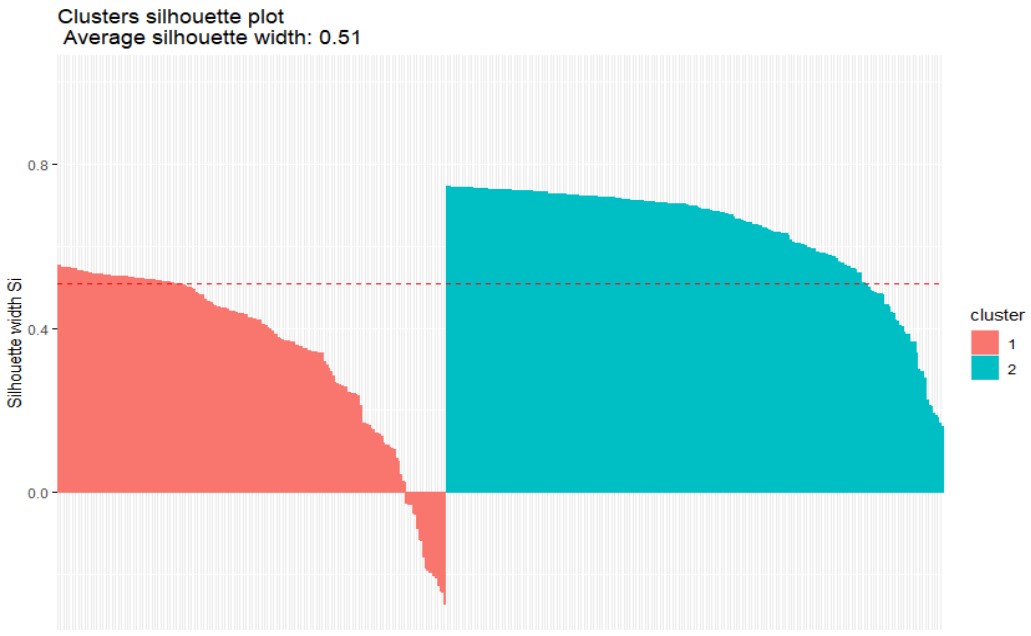


Figure 5.6 - Average Silhouette for *k*=2

Considering the second method, the *k* with the best classification is *k* =2, since it is the one with the highest value of average silhouette, revealing a better classification of objects when the dataset is split in two.

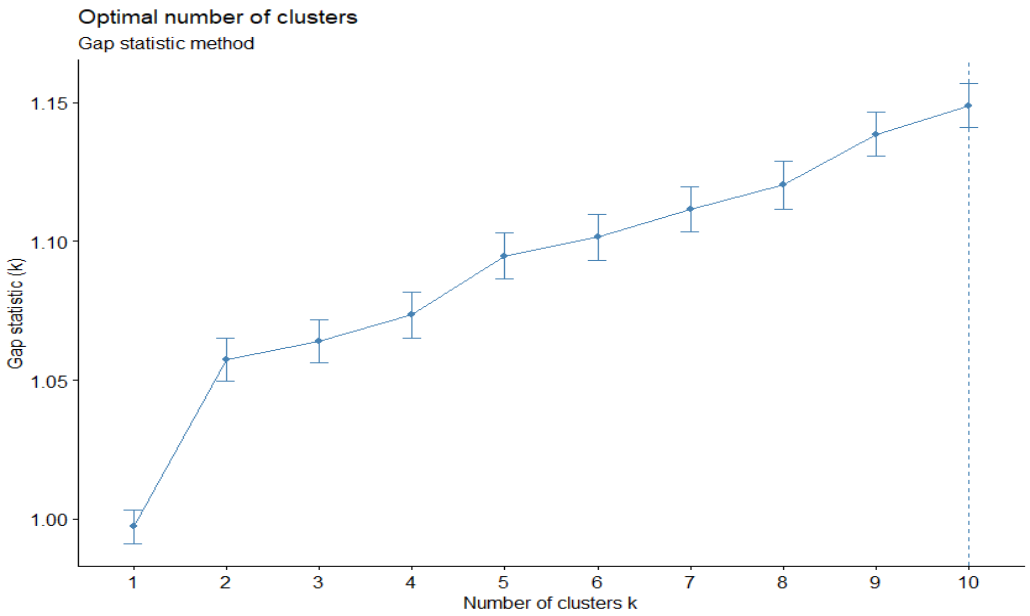


Figure 5.7 - Gap Statistic Method for K-Medoids

The Gap Statistic Method returns the following value of *k*=10, Figure 5.7. Since the variables under study are equal between the K-Means and the K-Medoids without the categorical variables, the application of the NbCLust package is equal to the one represented in Figure 5.4, with the ideal number of partitions being *k*=2. In conclusion, and after validation of the methods under study, the ideal number is *k*=2.

Assuming the value of *k*=2, the first cluster will be composed of 184 observations, which corresponds to 43.8% of the sample, while the second cluster will be composed of 161 observations, which corresponds to 56.2% of the sample.

K /Cluster Dimension	1	2	3	4	5	6	7	8
2	184	236						
3	171	123	126					
4	109	60	115	136				
5	95	69	115	24	117			
6	83	57	115	26	117	22		
7	83	57	66	26	72	94	22	
8	64	42	66	22	72	95	22	37

Table 5.12 - Absolute Frequency of Clusters with K-Medoids Algorithm (without categorical variables)

5.2.3. K-Medoids with categorical variables

The third group of variables under study introduces two categorical variables: Month and Period of the day. Since they are different variables from numerical variables, the validation process will be different to determine the ideal *k*. For this scenario, only the build and swap variation and the Average Silhouette will be considered.

K	Build and Swap	Variation (B&SK - B&SK-1)
2	5,42	-
3	5,07	-35,0%
4	4,91	-15,9%
5	4,76	-15,1%
6	4,63	-13,0%
7	4,51	-12,0%
8	4,44	-7,0%

Table 5.13 - K-Medoids Evaluation Process

K	Average silhouette width
2	0,18
3	0,09
4	0,05
5	0,06
6	0,05
7	0,04
8	0,04

Table 5.14 - Average Silhouette for K-Medoids

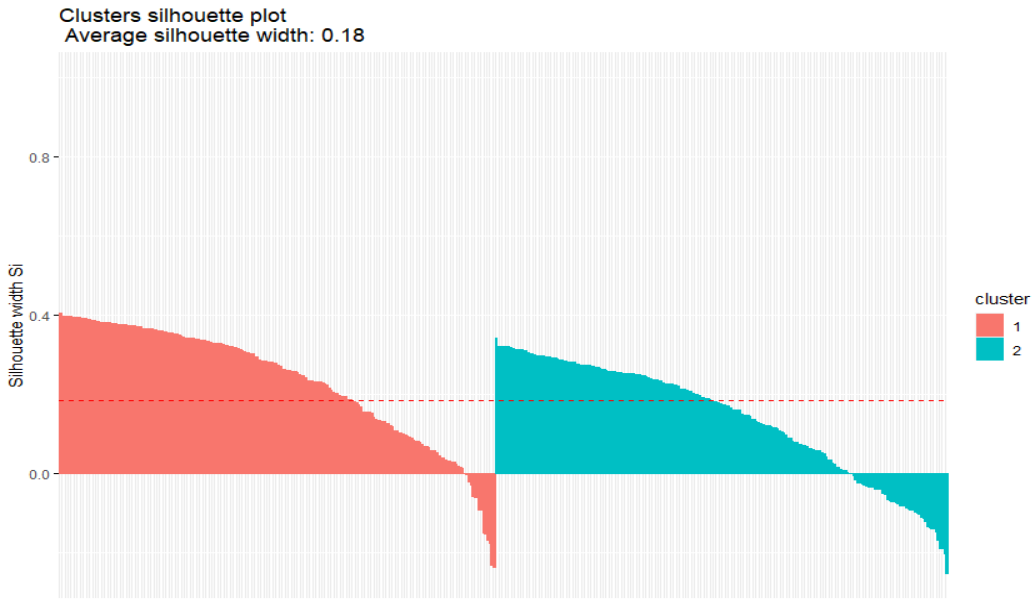


Figure 5.8 - Average Silhouette for *k*=2

Considering the two methods above, the ideal number of partitions is *k*=2. However, the Average Silhouette value is quite low, representing a low accuracy for the defined clusters.

Assuming the value of *k*=2, the first cluster will be composed of 206 observations, which corresponds to 49.0% of the sample, while the second cluster will be composed of 214 observations, which corresponds to 51.0% of the sample.

K /Cluster Dimension	1	2	3	4	5	6	7	8
2	206	214						
3	166	119	135					
4	74	114	99	133				
5	95	81	76	95	73			
6	86	73	62	77	60	62		
7	52	55	58	54	76	68	57	
8	43	53	58	48	73	66	57	22

Table 5.15 - Absolute Frequency of Clusters with K-Medoids Algorithm (with categorical variables)

5.3. DISCUSSION

In this topic, the output cluster from the tree methods above will be compared and interpreted. The final objective is to choose the method that produced the clusters with the best accuracy and the one that best explains pedestrian accidents in the city of Lisbon. Before proceeding with the analysis and interpretation of the clusters of each method, the clusters produced by K-Medoids without categorical variables is the one that produced a high Average Silhouette, revealing a better quality in the clusters produced.

5.3.1. K-Means with 2 clusters

Observing the graphic distribution of the clusters (Figure 5.9), it is possible to conclude that the center of the clusters is very close to each other, making neighboring objects very close. Cluster 2 presents a greater dispersion of objects in space, which is an indication of containing a greater number of outliers.

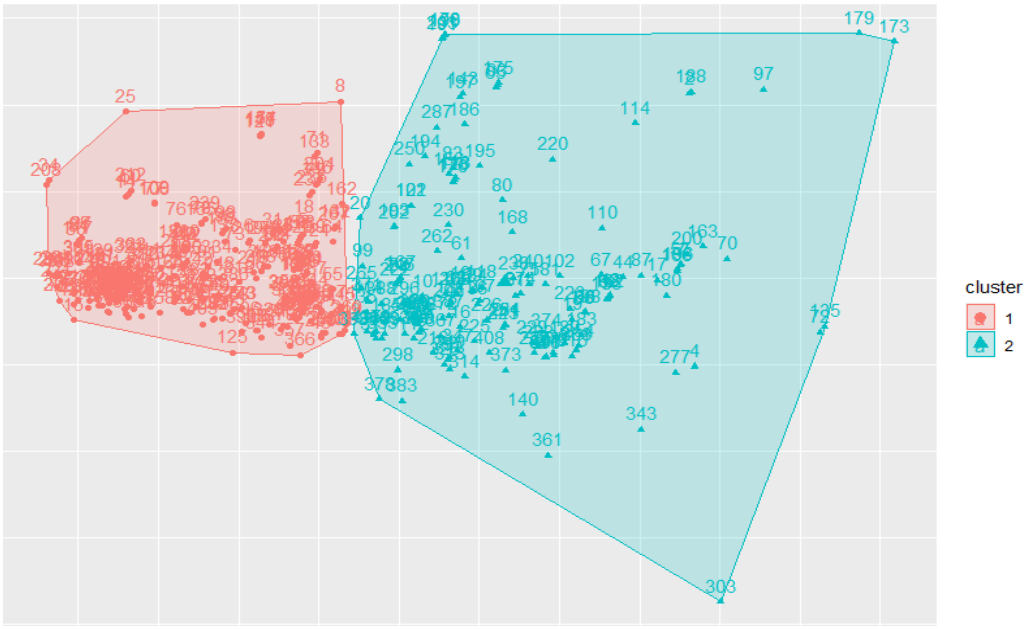


Figure 5.9 - Graphical representation of K-Means for *k*=2

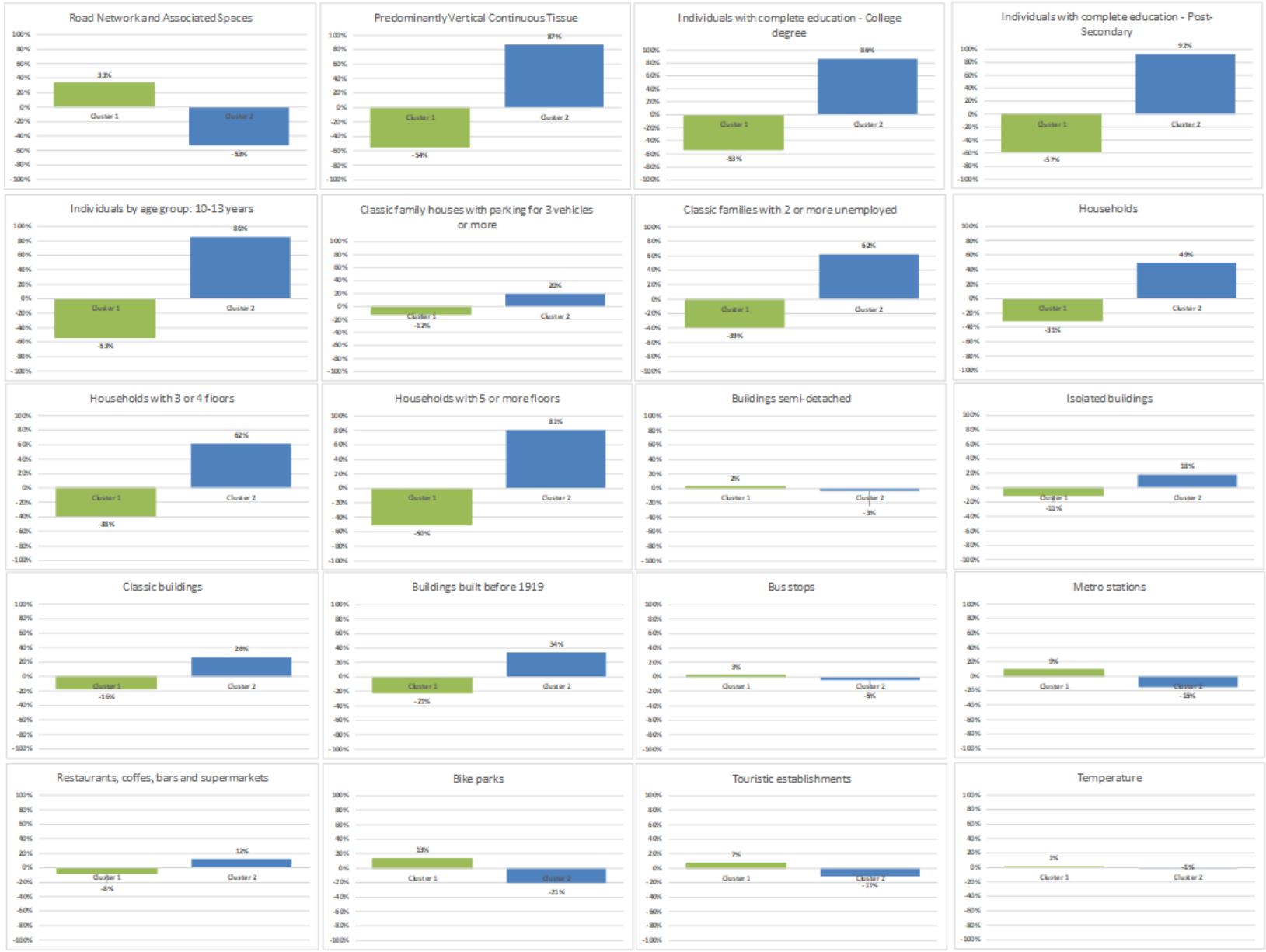


Figure 5.10 - Comparison of the winning variables in the K-Means cluster for *k*=2

To describe the clusters and their characteristics, the normalized values of the winning variables will be compared with the population mean (population mean is 0). A value greater than 0 indicates a positive correlation and a value below 0 indicates a negative correlation. The more the distance to the population means, the greater the importance of the variable under study. This comparison is performed using Figure 5.10. The 2 clusters can then be described as follows:

Cluster 1 - "Non-dense urban area": The first cluster contains 259 pedestrian accidents from the 420 records in the dataset, representing 61.6% of the sample.

This cluster is called a non-dense urban area, since it includes pedestrian accidents areas with the following characteristics: zones with a predominantly non-vertical building which indicates that the buildings are lower and follow a more horizontal arrangement. There is a lower density in terms of buildings. These are areas with a reasonable presence of road network and associated. In socio-economic terms, the inhabitants of these areas have a lower level of education, but unemployment rates are lower. Comparing cluster 1 with cluster 2, it is possible to see that the buildings are smaller. In terms of services and facilities, there is no clear distinction between the two groups. It should also be noted that the buildings in the zones of accidents with pedestrians in cluster 1 are slightly newer than those in cluster 2. In terms of temperature, and as previously mentioned, weather conditions do not play a decisive role in defining the location of the accident, as temperatures do not fluctuate on a large scale.

Cluster 2 - "Dense urban area": The second cluster contains 161 pedestrian accidents, representing 38.4% of the data.

This second cluster is called a dense urban area, since it has the following characteristics: Areas with a predominantly vertical building, which is an indicator that the buildings are taller and follow a more vertical layout. There is a higher density in terms of buildings. These are areas with a reduced network of roads and associates. In socioeconomic terms, these are areas where the inhabitants have a higher level of education (with a high incidence rate of graduates and those with post-secondary education), but unemployment rates are higher. Comparing cluster 1 with cluster 2, it is possible to see that the buildings are larger. In terms of services and facilities, there is no clear distinction between the two groups. It should also be noted that the buildings in the pedestrian accident zones in cluster 1 are slightly more recent than those in cluster 2. Another point to consider in cluster 2 buildings is that they follow a more classic layout, which indicates a greater propensity for housing areas (despite not being a variable with a great variation between both clusters).

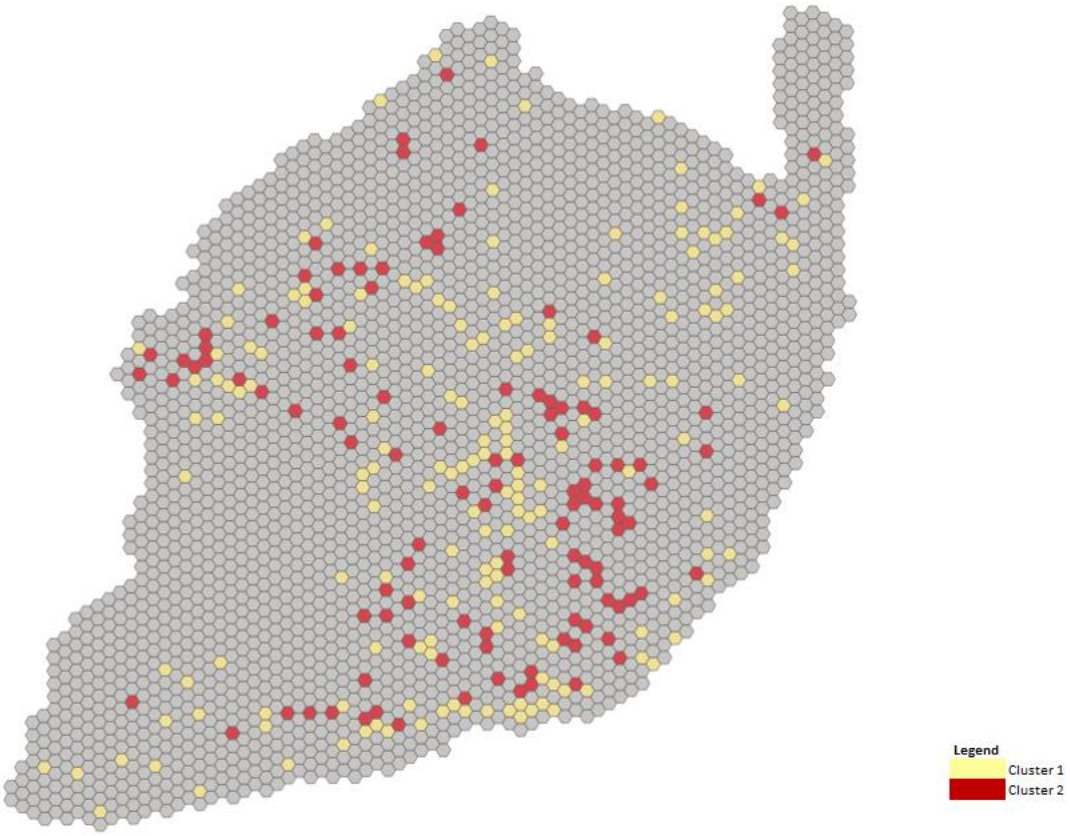


Figure 5.11 - Distribution of K-Means clusters on the hexagonal grid in the city of Lisbon

Considering the distribution of the two clusters across the map, shown in Figure 5.11, there is no geographical pattern to K-Means approach. The clusters are distributed throughout the city and are not grouped by zones or at the edges of the city of Lisbon. Cluster 1 is the most representative in this analysis, indicating that pedestrian accidents occur with greater incidence in areas with less urban characteristics in the city of Lisbon.

5.3.2. K-Medoids with 2 clusters (without categorical variables)

The distribution of points for K-Medoids without categorical variables, are quite similar to K-Means for $k = 2$, Figure 5.12. However, the Average Silhouette is higher, showing a more accurate classification.

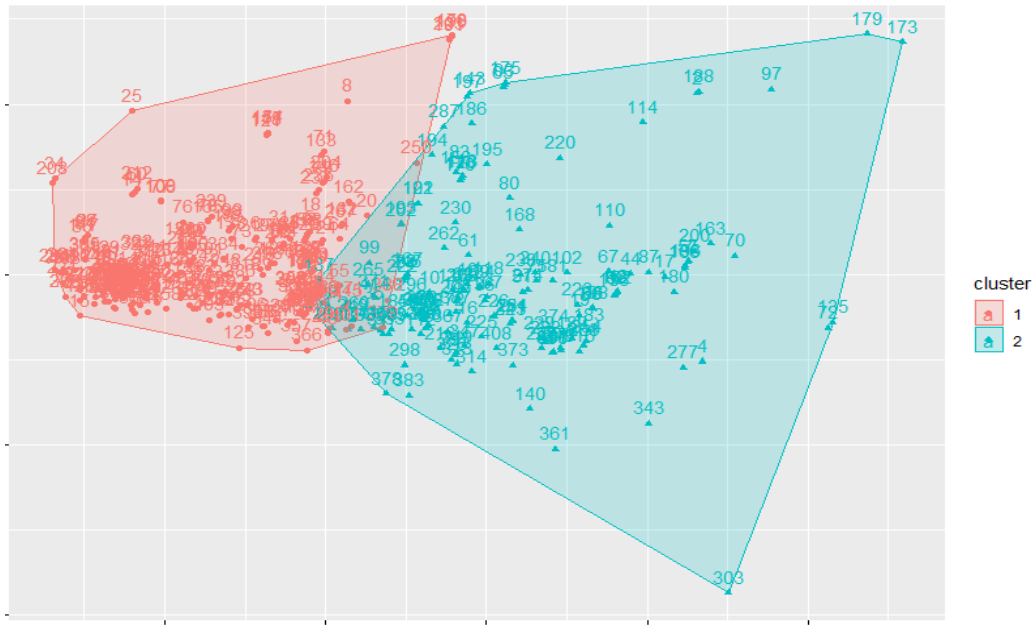


Figure 5.12 - Graphical representation of K-Medoids without categorical variables for *k*=2

Figure 5.13 - Comparison of the winning variables in the K-Medoids without categorical variables cluster for *k*=2

Based on Figure 5.13, the 2 clusters can then be described as follows:

Cluster 1 - "Non-dense urban area": The first cluster contains 184 pedestrian accidents from the 420 records in the dataset, representing 43.8% of the sample.

This cluster is called a non-dense urban area, since it includes pedestrian accidents areas with the following characteristics: zones with a predominantly non-vertical building which indicates that the buildings are lower and follow a more horizontal arrangement. There is a lower density in terms of buildings. These are areas with a reasonable presence of road network and associated. In socio-economic terms, the inhabitants of these areas have a lower level of education, but unemployment rates are lower. Comparing cluster 1 with cluster 2, it is possible to see that the buildings are smaller. In terms of services and facilities, there is no clear distinction between the two groups. It should also be noted that the buildings in the zones of accidents with pedestrians in cluster 1 are slightly newer than those in cluster 2. In terms of temperature, and as previously mentioned, weather conditions do not play a decisive role in defining the location of the accident, as temperatures do not fluctuate on a large scale. The cluster description is very similar with Cluster 1 from K-Means method.

Cluster 2 - "Dense urban area": The second cluster contains 236 pedestrian accidents, representing 56.2.% of the data.

This second cluster is called a dense urban area, since it has the following characteristics: Areas with a predominantly vertical building, which is an indicator that the buildings are taller and follow a more vertical layout. There is a higher density in terms of buildings. These are areas with a reduced network of roads and associates. In socioeconomic terms, these are areas where the inhabitants have a higher level of education (with a high incidence rate of graduates and those with post-secondary education), but unemployment rates are higher. Comparing cluster 1 with cluster 2, it is possible to see that the buildings are larger. In terms of services and facilities, there is no clear distinction between the two groups. It should also be noted that the buildings in the pedestrian accident zones in cluster 1 are slightly more recent than those in cluster 2. The cluster description is very similar with Cluster 2 from K-Means method.

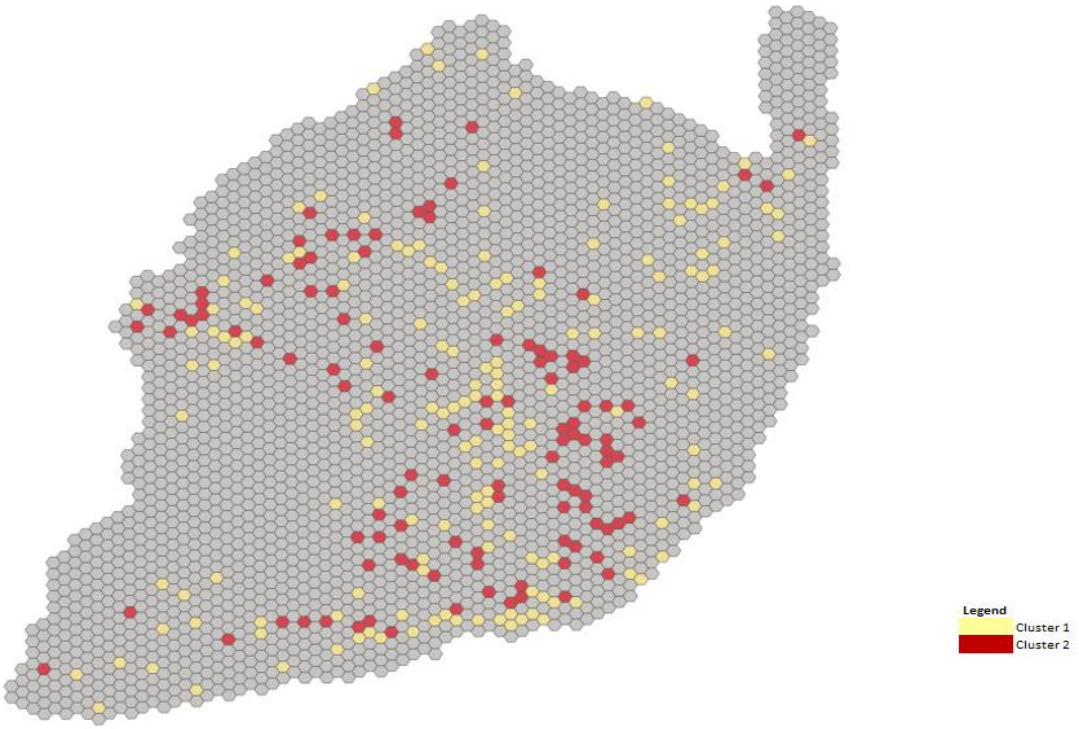


Figure 5.14 - Distribution of *k*=2 for K-Medoids clusters on the hexagonal grid in the city of Lisbon

Analyzing the distribution of the two clusters across the map, shown in Figure 5.14, there is no pattern. The clusters are distributed throughout the city and are not grouped by zones or at the edges of the city of Lisbon.

5.3.3. K-Medoids with 2 clusters (with categorical variables)

In this section the algorithm K-Medoids is used again. Although, this model considers the use of two categorical variables: Day Period and Month. Of the three methods, it was the one that registered the lowest Average Silhouette showing a poor classification for two partitions. Also, the distribution of points by cluster reveals an overlapping of clusters, contributing to the idea of the weak capacity of the model, Figure 5.15.

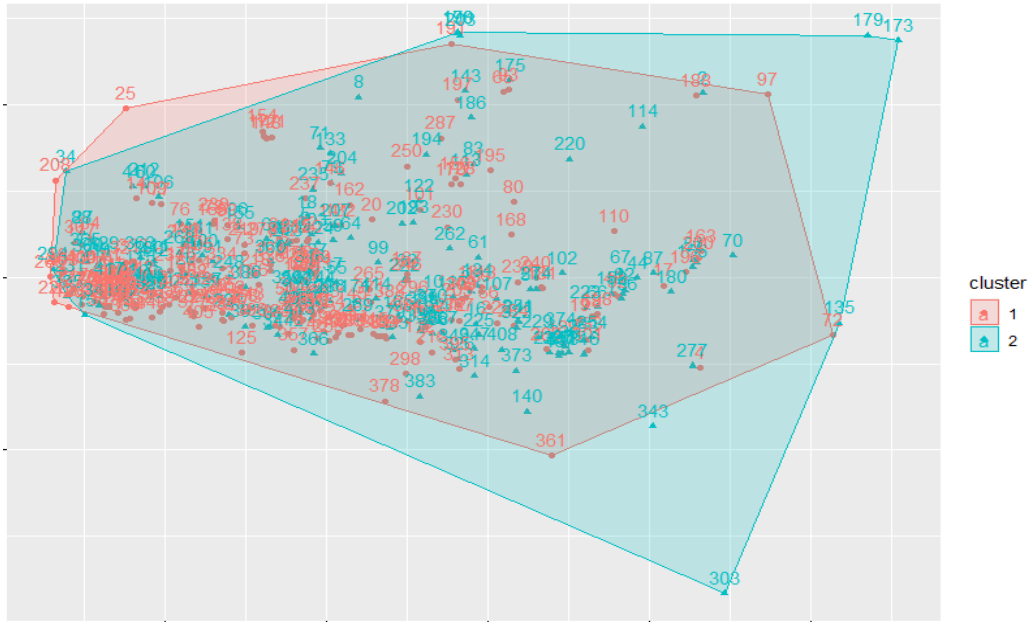
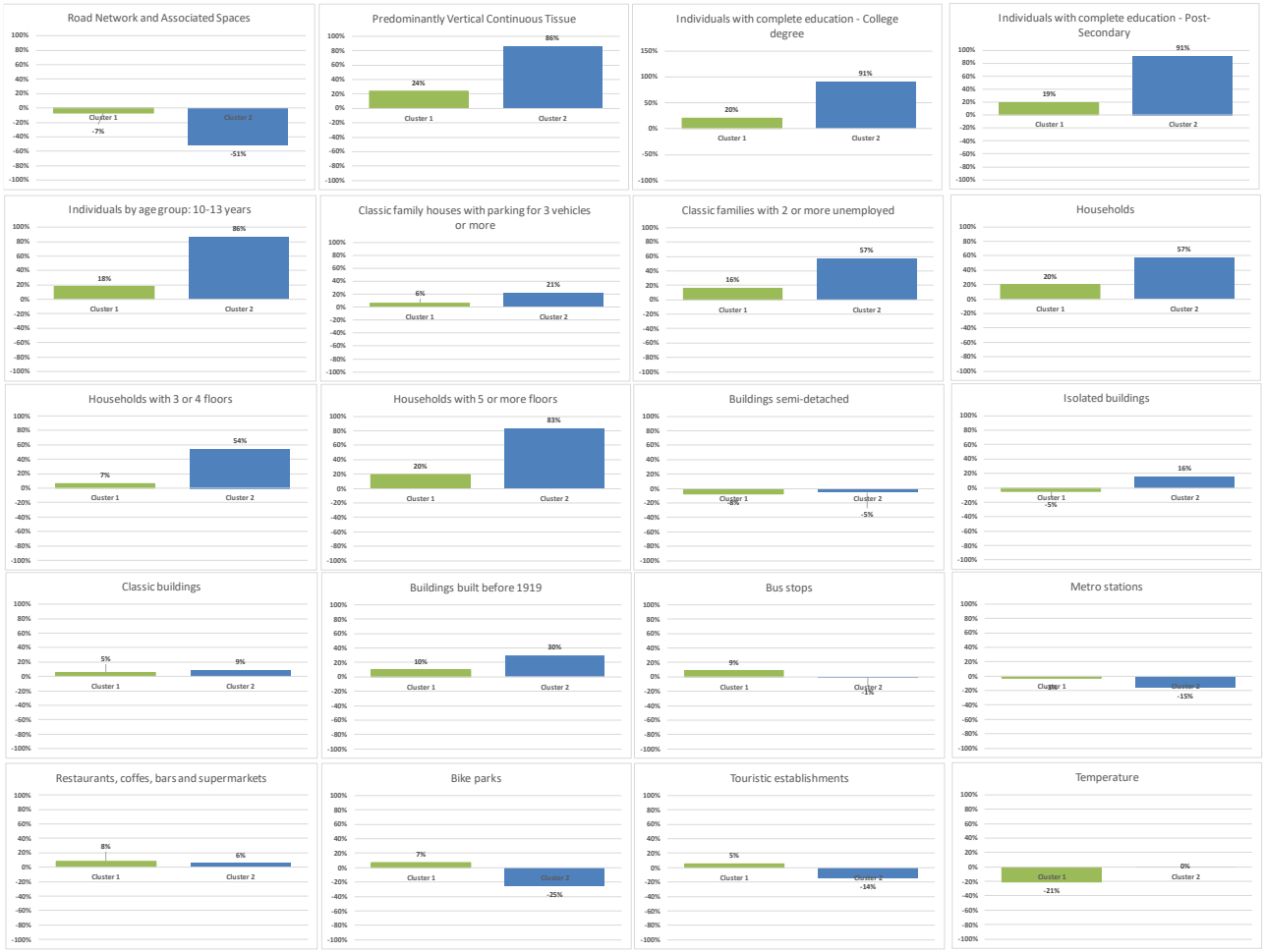


Figure 5.15 - Graphical representation of K-Medoids with categorical variables for *k*=2



Cluster 1

Month	#	Day Period	#
January	3	[0 - 3]	1
February	0	[4 - 7]	10
March	29	[8 - 11]	62
April	0	[12 - 15]	58
May	35	[16 - 19]	80
June	22	[20 - 23]	31
July	10		
August	0		
Septembe	48		
October	48		
Novembe	47		
Decembe	0		

Cluster 2

Month	#	Day Period	#
January	30	[0 - 3]	6
February	23	[4 - 7]	9
March	2	[8 - 11]	44
April	27	[12 - 15]	41
May	0	[16 - 19]	58
June	12	[20 - 23]	20
July	15		
August	32		
Septembe	0		
October	0		
Novembe	0		
Decembe	37		

Figure 5.16 - Comparison of the winning variables in the K-Medoids with categorical variables cluster for *k* = 2

Based on Figure 5.16, The 2 clusters can then be described as follows:

Cluster 1 - "Neutral Urban area": The first cluster contains 206 pedestrian accidents from the 420 records in the dataset, representing 49.0% of the sample.

This cluster is called neutral since the cluster values do not deviate from the average population. Analyzing its characteristics, it is possible to characterize the cluster area as follows: average number of road network and associates align with the average of the different areas of pedestrian accidents. Land areas with a more vertical arrangement, however it is not a determining characteristic of the cluster. In socioeconomic terms, education, and the number of unemployed are on average with the other areas. In terms of roadway characteristics and weather conditions the cluster does not deviate from the average. The cluster accidents occurred with greater incidence in Autumn, with the months of September, October and November registering a greater number of cases. The accidents occurred with greater incidence between 16-19 hours. The cluster is very similar to the data extracted from the data exploration in chapter 3. Data and Methods.

Cluster 2 - "Non-dense urban area.": The second cluster contains 214 pedestrian accidents, representing 51.0% of the data.

This cluster is called a non-dense urban area, since it has the following characteristics: zones with a predominantly vertical building, which indicates that the buildings are higher and follow a more vertical arrangement. Low incidence of road network and associated. In socioeconomic terms, highlight the presence of a population with a high academic level but at the same time a high preponderance of unemployed people. The areas in cluster 2 are characterized by the existence of household buildings, with a high number of floors. In terms of services and associates, cluster 2 does not differ from the population average. Temperature is not important for the cluster explanation. Analyzing the accident date, they occur with greater incidence in winter, with the months of December, January and February recording the highest numbers. These same accidents occurred with greater incidence between 16-19 hours.

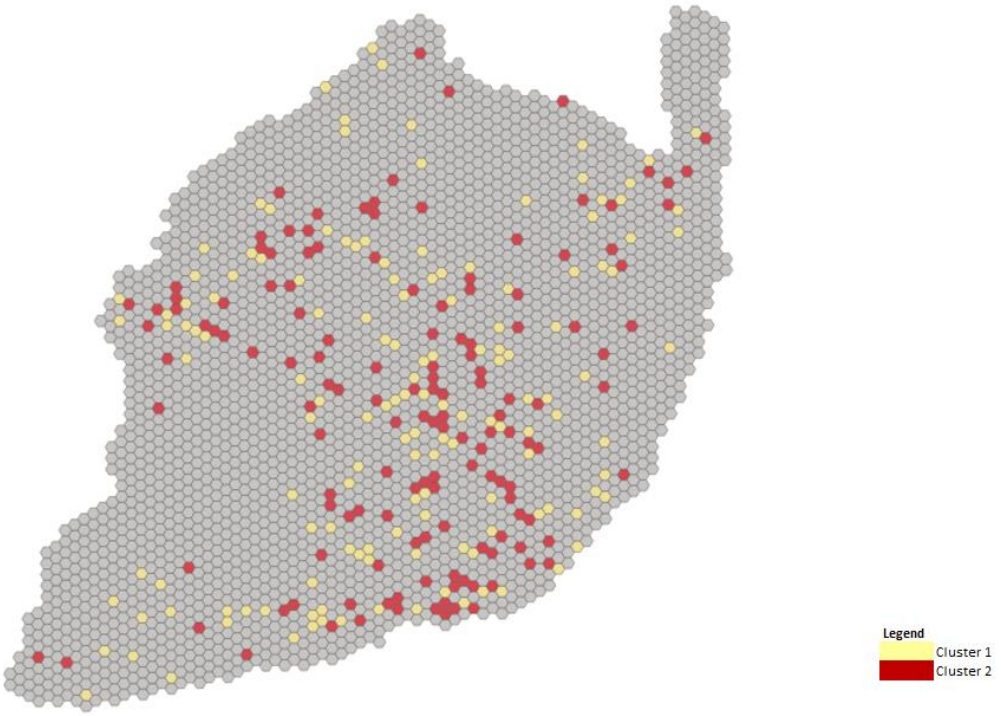


Figure 5.17 - Distribution of *k*=2 for K-Medoids clusters on the hexagonal grid in the city of Lisbon

Likewise for this method, there is no geographical pattern to the distribution of the two clusters obtained, Figure 5.17.

Comparing the three models estimated in this study and using Average silhouette width as a mechanism to compare the 3 outputs, the model estimated by K-Medoids without categorical variables is the winner, with a value of 0.51 for *k*=2. Second, the model estimated by K-Means with a value of 0.39 for *k*=2. Third, and with a less accurate result, is the K-Medoids with the inclusion of categorical variables, with a value of 0.18 for *k*=2. Although K-Medoids was the winning algorithm, the inclusion of categorical variables contributes to a poor quality in the association of observations to clusters, which is a bad principle and remove quality of analysis performed (Hawas and Guo 2019). Since the K-Medoids results with categorical variables were quite low, in this chapter the two outputs of K-Medoids and K-Means will be compared to understand the difference in the characteristics of both estimations and try to understand if there is a big difference in profiling the pedestrian accidents characteristics in Lisbon.

The winning model for K-Medoids create two clusters that allows to define pedestrian accidents in the city of Lisbon with the following characteristics:

Cluster 1 - Non-dense urban areas (43.8% of pedestrian accidents):

• Land Cover: Predominantly horizontal built-up areas (lower buildings); Presence of roads and associated spaces. Areas with fewer households.

• Socio-Economic characteristics: Low education level and low unemployment rates.

• Roadway characteristics: No distinction between clusters.

• Weather conditions: Not relevant.

Cluster 2 - Dense urban areas (56.2% of pedestrian accidents):

• Land Cover: Predominantly vertical built-up areas (lower buildings); Low presence of roads and associated spaces. Areas with the greatest number of households.

• Socio-Economic characteristics: High level of education and high unemployment rate.

• Roadway characteristics: No distinction between clusters.

• Weather conditions: Not relevant.

Analyzing the output for the winning model for K-Means, clusters can be defined as follows:

Cluster 1 - Non-dense urban areas (61.6% of pedestrian accidents):

• Land Cover: Predominantly horizontal built-up areas (lower buildings); Presence of roads and associated spaces. Areas with fewer households.

• Socio-Economic characteristics: Low education level and low unemployment rates.

• Roadway characteristics: No distinction between clusters.

• Weather conditions: Not relevant.

Cluster 2 - Dense urban areas (38.4% of pedestrian accidents):

• Land Cover: Predominantly vertical built-up areas (lower buildings); Low presence of roads and associated spaces. Areas with the greatest number of households.

• Socio-Economic characteristics: High level of education and high unemployment rate.

• Roadway characteristics: No distinction between clusters.

• Weather conditions: Not relevant.

The clusters generated by the K-Means and K-Medoids algorithm are very similar to each other. What really differs between them is the number of pedestrian accidents that constitute the clusters. For the K-Medoids algorithm, pedestrians' accidents in the city of Lisbon occur more frequently in dense and urban areas (56.2% of pedestrian accidents), for K-Means accidents are more common in less dense areas and areas with presence of roads and associates spaces (61.6% of pedestrian accidents).

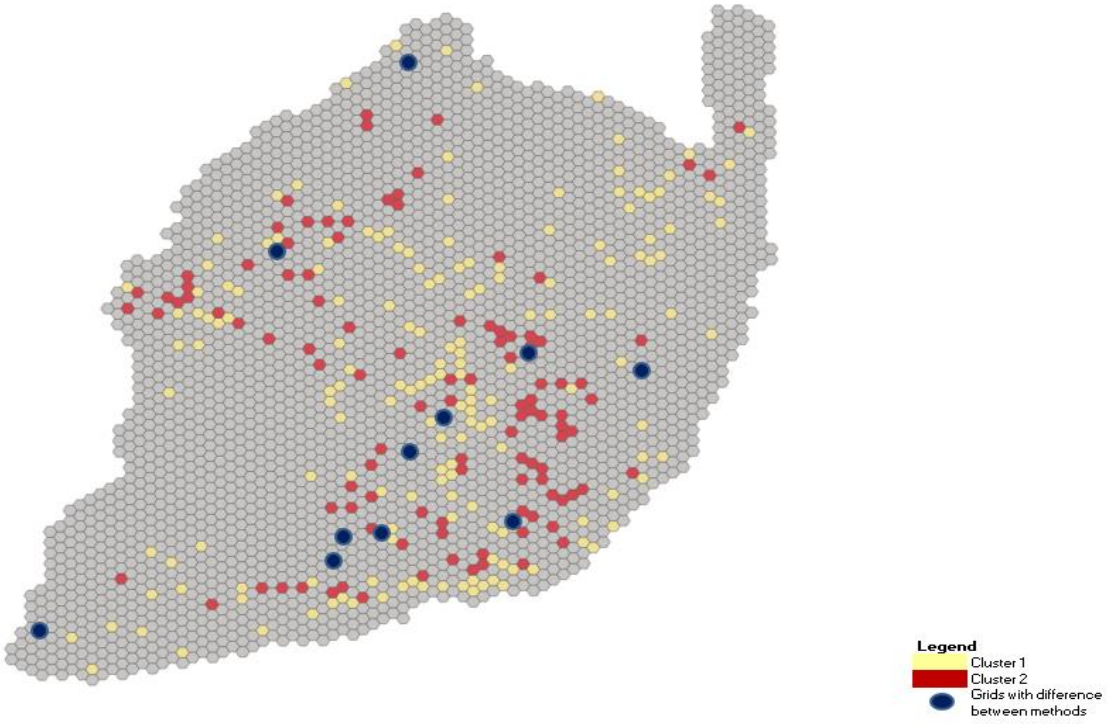


Figure 5.18 - Divergences for K-Means and K-Medoids clusters on the hexagonal grid in the city of Lisbon

Considering the map of the city of Lisbon, Figure 5.18, it is possible to observe 11 hexagonal grids of divergence between the two methods under study, and it is possible to conclude that the two algorithms worked similarly on the available dataset and that they extracted common patterns. As expected, K-Medoids assigned a different classification to some hexagonal grids, indicating that the accidents happen in denser urban areas.

However, despite the segmentation of the accident zones being quite similar, the winning model for this study is the K-Medoids, since it is the one that presents a better classification in the Average silhouette width.

6. CONCLUSIONS

This study presents an analysis of 420 accidents with pedestrians registered in the city of Lisbon, between 2013 and 2018, applying a cluster analysis to define and segment which are the most common factors in pedestrian accidents that have occurred. To conduct this study, a cluster analysis was performed using K-Means and K-Medoids algorithms, since they are a very strong tools to segment and identify patterns in datasets. The application of K-Medoids in this study was intended to test the ability to solve the problem of K-Means outliers, and the use of Gower Distance allows to use categorical variables under the models.

In conclusion of the study presented, it is possible to verify that pedestrian accidents in the city of Lisbon occur with greater preponderance in areas with more urban characteristics and with a higher density of buildings. This conclusion is in line with previous studies that indicate residential areas as areas with a higher probability of pedestrian accidents (D. J. Graham and Glaister 2003). However, it is important to note that most of the city of Lisbon follows a predominantly vertical distribution of buildings with an average area of 15 430 m^2. Road and associated networks are not abundant in these blackspots. Contrary to the studies mentioned in the literature review, pedestrian accidents in the city of Lisbon occur in areas where the population has a considerable level of education. The roadway characteristics and the weather conditions do not reveal a pattern different from the average population, which indicates that the pedestrian accidents occur with greater incidence in certain areas of the city (Table 3.3), with the end of summer and early autumn being a time with greater prevalence for the existence of accidents with pedestrians in the city of Lisbon (Figure 3.2). In the identified places, and to mitigate the risk of new accidents, it would be important to review the characteristics of the area (traffic signs, crosswalks, number of lanes, lane width, among others), to prevent future accidents, as well as sensitize people in the areas.

7. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

Considering the limitations of the study, the author finds four potential factors that could have added value to the analysis of the problem. Highlight first the lack of data for the characteristics of the victims. Being pointed out as a key topic in the literature review, the characteristics of the victim could have added value to the study, since we could measure and quantify which are the most common profiles in pedestrian accidents. They could define whether they would be men or women, young or old, Portuguese, or foreign, with basic or higher education, with or without the presence of alcohol, among other factors. The second factor to be considered in future studies, is related to roadway characteristics. It would be important to have data about the speed limit, the presence of crosswalks, the number of lanes on each road, the number of intersections, the length of each lane, among others. This information could describe the relevance of street furniture in pedestrian accidents. Another limiting factor of the study is related to the fact that the Censos is from 2011, and the study is being carried out with data collected more than 10 years ago. Finally, mention should be made of the lack of data on the economic incomes of the areas where pedestrian accidents were recorded. In previous studies, income has shown a strong relationship in the probability of pedestrian crashes, being considered a key factor.

As a suggestion for future segmentation studies on pedestrian accidents in urban areas, whether in Lisbon or in another city in the world, it would be important to consider the previously points to obtain more concrete and accurate results.

8. BIBLIOGRAPHY

Abdel-Aty, Mohamed, Jaeyoung Lee, Chowdhury Siddiqui, and Keechoo Choi. 2013. “Geographical Unit Based Analysis in the Context of Transportation Safety Planning.” Transportation Research Part A: Policy and Practice 49: 62–75. https://doi.org/10.1016/j.tra.2013.01.030.

Al-Ghamdi, Ali S. 2002. “Using Logistic Regression to Estimate the Influence of Accident Factors on Accident Severity.” Accident Analysis and Prevention 34 (6): 729–41. https://doi.org/10.1016/S0001-4575(01)00073-2.

Aljofey, Ali Moslah, and Khalil Alwagih. 2018. “Analysis of Accident Times for Highway Locations Using K-Means Clustering and Decision Rules Extracted from Decision Trees.” International Journal of Computer Applications Technology and Research 7 (1): 1–11. https://doi.org/10.7753/ijcatr0701.1001.

Anderson, Tessa K. 2009. “Kernel Density Estimation and K-Means Clustering to Profile Road Accident Hotspots.” Accident Analysis and Prevention 41 (3): 359–64. https://doi.org/10.1016/j.aap.2008.12.014.

Autoridade Nacional Segurança Rodoviária. 2018. “Sinistralidade Rodoviária.” 2018. http://www.ansr.pt/Estatisticas/RelatoriosDeSinistralidade/Pages/default.aspx.

Batool, Fatima, and Christian Hennig. 2021. “Clustering with the Average Silhouette Width.” EconPapers 158 (C).

Belhadi, Asma, Youcef Djenouri, Kjetil Nørvåg, Heri Ramampiaro, Florent Masseglia, and Jerry Chun Wei Lin. 2020. “Space–Time Series Clustering: Algorithms, Taxonomy, and Case Study on Urban Smart Cities.” Engineering Applications of Artificial Intelligence 95 (February): 103857. https://doi.org/10.1016/j.engappai.2020.103857.

Bernhoft, Inger Marie, and Gitte Carstensen. 2008. “Preferences and Behaviour of Pedestrians and Cyclists by Age and Gender.” Transportation Research Part F: Traffic Psychology and Behaviour 11 (2): 83–95. https://doi.org/10.1016/j.trf.2007.08.004.

Brown, Stephen, Kin Lo, and Thomas Lys. 1999. “Use of R-Squared in Accounting Research: Measuring Changes in Value Over the Last Four Decades” 28 (847).

Burton, D. 2010. “The Early History of Walkability.”

Cervero, Robert, Olga L. Sarmiento, Enrique Jacoby, Luis Fernando Gomez, and Andrea Neiman. 2009. “Influences of Built Environments on Walking and Cycling: Lessons from Bogotá.” International Journal of Sustainable Transportation 3 (4): 203–26. https://doi.org/10.1080/15568310802178314.

Clifton, K.J, K Fults, Burnier, C., and Kreamer Fults, K. 2004. “Women’s Involvement in Pedestrian–Vehicle Crashes. Influence of Personal and Environmental Factors.” Transportation Research Board.

D. Magnusson, and L.R. Bergman. 2001. International Encyclopedia of the Social & Behavioral Sciences.

Dameri, Renata Paola. 2014. “Council for Innovative Research.” Journal of Advances in Chemistry 10 (1): 2146–61. https://doi.org/10.13140/RG.2.1.3973.9042.

Damsere-Derry, James, Beth E. Ebel, Charles N. Mock, Francis Afukaar, and Peter Donkor. 2010. “Pedestrians’ Injury Patterns in Ghana.” Accident Analysis and Prevention 42 (4): 1080–88. https://doi.org/10.1016/j.aap.2009.12.016.

Das, Subasish, Anandi Dutta, Raul Avelar, Karen Dixon, Xiaoduan Sun, and Mohammad Jalayer. 2019. “Supervised Association Rules Mining on Pedestrian Crashes in Urban Areas: Identifying Patterns for Appropriate Countermeasures.” International Journal of Urban Sciences 23 (1): 30–48. https://doi.org/10.1080/12265934.2018.1431146.

Demetriades, Demetrios, James Murray, Matthew Martin, George Velmahos, Ali Salim, Kathy Alo, and Peter Rhee. 2004. “Pedestrians Injured by Automobiles: Relationship of Age to Injury Type and Severity.” Journal of the American College of Surgeons 199 (3): 382–87. https://doi.org/10.1016/j.jamcollsurg.2004.03.027.

Dong, Ni, Helai Huang, Pengpeng Xu, Zhuodi Ding, and Duo Wang. 2014. “Evaluating Spatial-Proximity Structures in Crash Prediction Models at the Level of Traffic Analysis Zones.” Transportation Research Record: Journal of the Transportation Research Board 2432 (1): 46–52. https://doi.org/10.3141/2432-06.

Dumbaugh, Eric, and Wenhao Li. 2011. “Designing for the Safety of Pedestrians, Cyclists, and Motorists in Urban Environments.” Journal of the American Planning Association 77 (1): 69–88. https://doi.org/10.1080/01944363.2011.536101.

European Commission. 2018. “Traffic Safety Basic Facts on Pedestrians.” European Road Safety Observatory, 1–24. https://ec.europa.eu/transport/road_safety/sites/roadsafety/files/pdf/statistics/dacota/bfs2018_pedestrians.pdf.

Firoz, Mohammed, and Vinod Kumar. 2017. Chapter 16: Transforming Economy of Calicut to Smart Economy. Smart Economy in Smart Cities. https://doi.org/10.1007/978-981-10-1610-3.

Forthofer, Ronald N., Eun Sul Lee, and Mike Hernandez. 2007. Biostatistics.

Fountas, Grigorios, Achille Fonzone, Niaz Gharavi, and Tom Rye. 2020. “The Joint Effect of Weather and Lighting Conditions on Injury Severities of Single-Vehicle Accidents.” Analytic Methods in Accident Research 27: 100124. https://doi.org/10.1016/j.amar.2020.100124.

Fylan, F., A. Hughes, J. M. Wood, and D. B. Elliott. 2018. “Why Do People Drive When They Can’t See Clearly?” Transportation Research Part F: Traffic Psychology and Behaviour 56: 123–33. https://doi.org/10.1016/j.trf.2018.04.005.

Giffinger, Rudolf, Christian Fertner, Hans Kramar, and Evert Meijers. 2007. “City-Ranking of European Medium-Sized Cities.” Centre of Regional Science, Vienna UT, no. October.

Graham, Daniel, Stephen Glaister, and Richard Anderson. 2005. “The Effects of Area Deprivation on the Incidence of Child and Adult Pedestrian Casualties in England.” Accident Analysis and Prevention 37 (1): 125–35. https://doi.org/10.1016/j.aap.2004.07.002.

Graham, Daniel J., and Stephen Glaister. 2003. “Spatial Variation in Road Pedestrian Casualties: The Role of Urban Scale, Density and Land-Use Mix.” Urban Studies 40 (8): 1591–1607. https://doi.org/10.1080/0042098032000094441.

Hadayeghi, Alireza, Amer S. Shalaby, and Bhagwant N. Persaud. 2010. “Development of Planning Level Transportation Safety Tools Using Geographically Weighted Poisson Regression.” Accident Analysis and Prevention 42 (2): 676–88. https://doi.org/10.1016/j.aap.2009.10.016.

Harrison, C., B. Eckman, R. Hamilton, P. Hartswick, J. Kalagnanam, J. Paraszczak, and P. Williams. 2010. “Foundations for Smarter Cities.” IBM Journal of Research and Development 54 (4): 1–16. https://doi.org/10.1147/JRD.2010.2048257.

Hawas, Ahmed Refaat, and Yanhui Guo. 2019. Neutrosophic Set in Medical Image Analysis.

Hebert Martinez, Kristie L., and Bryan E. Porter. 2004. “The Likelihood of Becoming a Pedestrian Fatality and Drivers’ Knowledge of Pedestrian Rights and Responsibilities in the Commonwealth of Virginia.” Transportation Research Part F: Traffic Psychology and Behaviour 7 (1): 43–58. https://doi.org/10.1016/j.trf.2003.11.001.

Huang, Helai, and Hong Chor Chin. 2010. “Modeling Road Traffic Crashes with Zero-Inflation and Site-Specific Random Effects.” Statistical Methods and Applications 19 (3): 445–62. https://doi.org/10.1007/s10260-010-0136-x.

INE - Instituto Nacional de Estatística. 2018. Mobilidade e Funcionalidade Do Território Nas Áreas Metropolitanas Do Porto e de Lisboa : 2017. https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_publicacoes&PUBLICACOESpub_boui=349495406&PUBLICACOESmodo=2.

Instituto Nacional de Estatística. 2018. “População Residente (N.o) Por Local de Residência (NUTS - 2013).” 2018. https://www.ine.gob.gt/ine/.

Instituto Português do Mar e Atmosfera. 2020. “CLASSIFICAÇÃO DO VENTO.” 2020.

Jacobsen, Peter Lyndon. 2015. “Safety in Numbers: More Walkers and Bicyclists, Safer Walking and Bicycling.” Injury Prevention 21 (4): 271–75. https://doi.org/10.1136/ip.9.3.205rep.

Johnson, Emily, Judy Geyer, David R Ragland, and Nirmeet Rai. 2004. “Low Income Childhood Pedestrian Injury: Understanding the Disparate Risk.”

Julien Boccard, and Serge Rudaz. 2013. Proteomic and Metabolomic Approaches to Biomarker Discovery.

Kim, Karl, I. Made Brunner, and Eric Y. Yamashita. 2006. “Influence of Land Use, Population, Employment, and Economic Activity on Accidents.” Transportation Research Record, no. 1953: 56–64. https://doi.org/10.3141/1953-07.

Kim, Karl, Pradip Pant, and Eric Yamashita. 2010. “Accidents and Accessibility: Measuring Influences of Demographic and Land Use Variables in Honolulu, Hawaii.” Transportation Research Record, no. 2147: 9–17. https://doi.org/10.3141/2147-02.

Kim, Karl, and Eric Y. Yamashita. 2007. “Using a K-Means Clustering Algorithm to Examine Patterns of Pedestrian Involved Crashes in Honolulu, Hawaii.” Journal of Advanced Transportation 41 (1): 69–89. https://doi.org/10.1002/atr.5670410106.

Kinga Ivan, Ionel Haidu, József Benedek, and Silviu Marian Ciobanu. 2015. “Identification of Traffic Accident Risk-Prone Areas under Low Lighting Conditions.” Natural Hazards and Earth System Sciences 15(9), 2015.

Kjell Johnson, and Max Kuhn. 2013. Applied Predictive Modeling.

Kumar, Sachin, and Durga Toshniwal. 2015. “A Data Mining Framework to Analyze Road Accident Data.” Journal of Big Data 2 (1). https://doi.org/10.1186/s40537-015-0035-y.

LaScala, Elizabeth A., Paul J. Gruenewald, and Fred W. Johnson. 2004. “An Ecological Study of the Locations of Schools and Child Pedestrian Injury Collisions.” Accident Analysis and Prevention 36 (4): 569–76. https://doi.org/10.1016/S0001-4575(03)00063-0.

Lee, Chris, and Mohamed Abdel-Aty. 2005. “Comprehensive Analysis of Vehicle-Pedestrian Crashes at Intersections in Florida.” Accident Analysis and Prevention 37 (4): 775–86. https://doi.org/10.1016/j.aap.2005.03.019.

Lee, Jaeyoung, Mohamed Abdel-Aty, and Ximiao Jiang. 2014. “Development of Zone System for Macro-Level Traffic Safety Analysis.” Journal of Transport Geography 38: 13–21. https://doi.org/10.1016/j.jtrangeo.2014.04.018.

Levine, Ned, Karl E. Kim, and Lawrence H. Nitz. 1995. “Spatial Analysis of Honolulu Motor Vehicle Crashes: I. Spatial Patterns.” Accident Analysis and Prevention 27 (5): 663–74. https://doi.org/10.1016/0001-4575(95)00017-T.

Li, Yue, and Geoff Fernie. 2010. “Pedestrian Behavior and Safety on a Two-Stage Crossing with a Center Refuge Island and the Effect of Winter Weather on Pedestrian Compliance Rate.” Accident Analysis and Prevention 42 (4): 1156–63. https://doi.org/10.1016/j.aap.2010.01.004.

Loukaitou-Sideris, Anastasia, Robin Liggett, and Hyun Gun Sung. 2007. “Death on the Crosswalk: A Study of Pedestrian-Automobile Collisions in Los Angeles.” Journal of Planning Education and Research 26 (3): 338–51. https://doi.org/10.1177/0739456X06297008.

Marcello D’Orazio. 2013. “Distances with Mixed-Type Variables, Some Modified Gower’s Coefficients.” Journal of Chemical Information and Modeling 53 (9): 1689–99.

Maze, Thomas H, Christian Sax, Neal Hawkins, Ames Iowa State University, Transportation Iowa Department of, and Consortium Midwest Transportation. 2008. “Clear Zone – A Synthesis of Practice and an Evaluation of the Benefits of Meeting the 10-Ft Clear Zone Goal on Urban Streets,” no. November 2008: 158p. http://www.ctre.iastate.edu/reports/clear_zone_report.pdf https://trid.trb.org/view/878049.

Miškinis, Paulius, and Vaida Valuntaite. 2011. “Mathematical Simulation of the Correlation between the Frequency of Road Traffic Accidents and Driving Experience.” DOAJ.

Morency, P., and M. S. Cloutier. 2006. “From Targeted ‘Black Spots’ to Area-Wide Pedestrian Safety.” Injury Prevention 12 (6): 360–64. https://doi.org/10.1136/ip.2006.013326.

Moudon, Anne Vernez, Lin Lin, Junfeng Jiao, Philip Hurvitz, and Paula Reeves. 2011. “The Risk of Pedestrian Injury and Fatality in Collisions with Motor Vehicles, a Social Ecological Study of State Routes and City Streets in King County, Washington.” Accident Analysis and Prevention 43 (1): 11–24. https://doi.org/10.1016/j.aap.2009.12.008.

Naik, Bhaven, Li Wei Tung, Shanshan Zhao, and Aemal J. Khattak. 2016. “Weather Impacts on Single-Vehicle Truck Crash Injury Severity.” Journal of Safety Research 58: 57–65. https://doi.org/10.1016/j.jsr.2016.06.005.

Nations, United. 2018. “The World ’s Cities in 2018.” The World’s Cities in 2018 - Data Booklet (ST/ESA/ SER.A/417), 34.

Olszewski, Piotr, Piotr Szagała, Maclej Wolański, and Anna Zielińska. 2015. “Pedestrian Fatality Risk in Accidents at Unsignalized Zebra Crosswalks in Poland.” Accident Analysis and Prevention 84: 83–91. https://doi.org/10.1016/j.aap.2015.08.008.

Ossenbruggen, Paul J., Jyothi Pendharkar, and John Ivan. 2001. “Roadway Safety in Rural and Small Urbanized Areas.” Accident Analysis and Prevention 33 (4): 485–98. https://doi.org/10.1016/S0001-4575(00)00062-2.

Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. 2004. Introduction to Data Mining.

Park, Hae Sang, and Chi Hyuck Jun. 2009. “A Simple and Fast Algorithm for K-Medoids Clustering.” Expert Systems with Applications. Vol. 36. Elsevier Ltd. https://doi.org/10.1016/j.eswa.2008.01.039.

Pfeffer, K., H. P. Fagbemi, and S. Stennet. 2010. “Adult Pedestrian Behavior When Accompanying Children on the Route to School.” Traffic Injury Prevention 11 (2): 188–93. https://doi.org/10.1080/15389580903548576.

Prato, Carlo Giacomo, Victoria Gitelman, and Shlomo Bekhor. 2012. “Mapping Patterns of Pedestrian Fatal Accidents in Israel.” Accident Analysis and Prevention 44 (1): 56–62. https://doi.org/10.1016/j.aap.2010.12.022.

Preusser, David F, Joann K Wells, Allan F Williams, and Helen B Weinstein. 2002. “Pedestrian Crashes in Washington, DC and Baltimore.” PubMed.

Pulugurtha, Srinivas S., Venkata Ramana Duddu, and Yashaswi Kotagiri. 2013. “Traffic Analysis Zone Level Crash Estimation Models Based on Land Use Characteristics.” Accident Analysis and Prevention 50: 678–87. https://doi.org/10.1016/j.aap.2012.06.016.

Reed, Randal, and Siddhartha Sen. 2005. “Racial Differences and Pedestrian Safety: Some Evidence from Maryland and Implications for Policy.” Journal of Public Transportation 8 (2): 37–61. https://doi.org/10.5038/2375-0901.8.2.3.

Richard A Retting, Susan A Ferguson, and Anne T McCartt. 2003. “A Review of Evidence-Based Traffic Engineering Measures Designed to Reduce Pedestrian-Motor Vehicle Crashes.”

Ryb, Gabriel E., Patricia C. Dischinger, Joseph A. Kufera, and Carl A. Soderstrom. 2007. “Social, Behavioral and Driving Characteristics of Injured Pedestrians: A Comparison with Other Unintentional Trauma Patients.” Accident Analysis and Prevention 39 (2): 313–18. https://doi.org/10.1016/j.aap.2006.08.004.

Sivasankaran, Sathish Kumar, and Venkatesh Balasubramanian. 2020. “Exploring the Severity of Bicycle – Vehicle Crashes Using Latent Class Clustering Approach in India.” Journal of Safety Research 72 (December): 127–38. https://doi.org/10.1016/j.jsr.2019.12.012.

Sterling, Thomas, Matthew Anderson, and Maciej Brodowicz. 2018. High Performance Computing - Modern Systems and Practices.

Steven D. Brown, Romá Tauler, and Beata Walczak. 2009. Comprehensive Chemometrics - Chemical and Biochemical Data Analysis.

Su, Kehua, Jie Li, and Hongbo Fu. 2011. “Smart City and the Applications.” 2011 International Conference on Electronics, Communications and Control, ICECC 2011 - Proceedings, 1028–31. https://doi.org/10.1109/ICECC.2011.6066743.

Sullman, Mark J.M., Abigail Thomas, and Amanda N. Stephens. 2012. “The Road User Behaviour of School Students in Belgium.” Accident Analysis and Prevention 48: 495–504. https://doi.org/10.1016/j.aap.2012.03.004.

Sun, Ming, Xiaoduan Sun, and Donghui Shan. 2019. “Pedestrian Crash Analysis with Latent Class Clustering Method.” Accident Analysis and Prevention 124 (June 2018): 50–57. https://doi.org/10.1016/j.aap.2018.12.016.

Sze, N. N., and S. C. Wong. 2007. “Diagnostic Analysis of the Logistic Model for Pedestrian Injury Severity in Traffic Crashes.” Accident Analysis and Prevention 39 (6): 1267–78. https://doi.org/10.1016/j.aap.2007.03.017.

Tay, Richard, Jaisung Choi, Lina Kattan, and Amjad Khan. 2011. “A Multinomial Logit Model of Pedestrian-Vehicle Crash Severity.” International Journal of Sustainable Transportation 5 (4): 233–49. https://doi.org/10.1080/15568318.2010.497547.

Tay, Richard, and Shakil Mohammad Rifaat. 2007. “Factors Contributing to the Severity of Intersection Crashes.” Journal of Advanced Transportation 41 (3): 245–65. https://doi.org/10.1002/atr.5670410303.

Theofilatos, Athanasios, and Dimitrios Efthymiou. 2012. “Investigation of Pedestrians’ Accident Patterns in Greater Athens Area.” Procedia - Social and Behavioral Sciences 48: 1897–1906. https://doi.org/10.1016/j.sbspro.2012.06.1164.

Tibshirani, Robert, Guenther Walther, and Trevor Hastie. 2001. “Estimating the Number of Data Clusters via the Gap Statistic.” Journal of the Royal Statistical Society: Series B.

Whitelegg, J. 1987. “A Geography of Road Traffic Accidents.” Transactions - Institute of British Geographers 12 (2): 161–76. https://doi.org/10.2307/622525.

WHO, WHO World Heatlh Organization. 2018. “GLOBAL STATUS REPORT ON ROAD SAFETY.”

Yueying Wang, Md. Mazharul Haque, Hoong Chor Chin, and Jelphine Goh Jie Yun. 2013. “Injury Severity of Pedestrian Crashes in Singapore.”

Zeng, Qiang, and Helai Huang. 2014. “Bayesian Spatial Joint Modeling of Traffic Crashes on an Urban Road Network.” Accident Analysis and Prevention 67: 105–12. https://doi.org/10.1016/j.aap.2014.02.018.

Zhang, Lemin, Ruoxi Zhang, and Biao Yin. 2021. “The Impact of the Built-up Environment of Streets on Pedestrian Activities in the Historical Area.” Alexandria Engineering Journal 60 (1): 285–300. https://doi.org/10.1016/j.aej.2020.08.008.

9. APPENDIX

Group	Variables
Land Cover	Avg. Predominantly Vertical Continuous Tissue
	Avg. Predominantly Horizontal Continuous Tissue
	Avg. Discontinuous Built-Up Tissue
	Avg. Sparse discontinuous built fabric
	Avg. Parking Areas
	Avg. Empty Places
	Avg. Industries
	Avg. Commercial Areas
	Avg. Agricultural Facilities
	Avg. Non-renewable energy production infrastructures
	Avg. Waste and water treatment infrastructures
	Avg. Road Network and Associated Spaces
	Avg. Railway Network and Associated Spaces
	Avg. Port Terminals
	Avg. Marinas and fishing docks
	Avg. Airports
	Avg. Stone Quarry
	Avg. Landfills
	Avg. Trash and Scrap
	Avg. Areas under construction
	Avg. Golf Courses
	Avg. Sports Facilities
	Avg. Leisure Equipment
	Avg. Cultural Facilities
	Avg. Tourist Facilities
	Avg. Parks Gardens
	Avg. Cemeteries
	Avg. Other tourist equipment's and facilities
	Avg. Vineyards
	Avg. Orchards
	Avg. Olive Tree
	Avg. Temporary cultures
	Avg. Complex cultural and parcel mosaics
	Avg. Agriculture with natural and semi-natural spaces
	Avg. Improved pastures
	Avg. Spontaneous Pastures
	Avg. Eucalyptus Forests
	Avg. Forests of other hardwoods
	Avg. Pinus pinaster forests
	Avg. Stone Pine Forests
	Avg. Forests of other coniferous trees
	Avg. Bushes
	Avg. Beaches, dunes and coastal sand

	Avg. Salt marshes
	Avg. Natural watercourses
	Avg. Artificial lakes and ponds
	Avg. River mouths

Table 9.1 - List of variables of the "Land Cover" group

```
Recursive feature selection

Outer resampling method: Cross-Validated (10 fold, repeated 5 times)

Resampling performance over subset size:

 Variables   RMSE Rsquared    MAE  RMSESD RsquaredSD   MAESD Selected
         1 0.7023   0.5217 0.4763 0.14789     0.1908 0.08829
         2 0.5147   0.7401 0.4096 0.08033     0.1087 0.05979        *
         3 0.5997   0.6831 0.5013 0.05689     0.1108 0.04638
         4 0.6311   0.6706 0.5253 0.05283     0.1091 0.04212
         5 0.6677   0.6408 0.5522 0.05078     0.1174 0.04279
         6 0.5343   0.7481 0.4396 0.06834     0.1049 0.04672
         7 0.5649   0.7259 0.4683 0.06207     0.1133 0.04462
```

Figure 9.1 – RFE output for the "Land Cover" group variables

Group	**Variables**
Socio-Economic	Avg. Number of classic families with 2 or more unemployed
	Avg. Number of classic families with 1 unemployed
	Avg. Number of classic families
	Avg. Number of classic families with no unemployed
	Avg. Number of institutional families
	Avg. Number of families with 1 or 2 people
	Avg. Number of families with 3 or 4 people
	Avg. Number of families with people under 15 years
	Avg. Number of families with people over 65 years
	Avg. Number of institutional families
	Avg. Number of unemployed individuals
	Avg. Number of individuals employed - primary sector
	Avg. Number of individuals employed - secondary sector
	Avg. Number of individuals employed - tertiary sector
	Avg. Number of individuals employed
	Number of resident individuals studying in the municipality of residence
	Avg. Number of pensioners and retirees
	Avg. Number of individuals without economic activity
	Number of resident individuals working in the municipality of residence
	Avg. Number of individuals with complete education - 1st cycle
	Avg. Number of individuals with complete education - 2nd cycle
	Avg. Number of individuals with complete education - 3rd cycle
	Avg. Number of individuals with complete education - post-Secondary
	Avg. Number of individuals with complete education - Secondary
	Avg. Number of individuals with complete education - College degree

	Avg. Number of individuals attending with complete education - 1st cycle
	Avg. Number of individuals attending with complete education - 2nd cycle
	Avg. Number of individuals attending with complete education - 3rd cycle
	Avg. Number of individuals attending with complete education - post-Secondary
	Avg. Number of individuals attending with complete education - Secondary
	Avg. Number of individuals attending with complete education - College degree
	Avg. Number of illiterate individuals
	Avg. Number of individuals presents
	Avg. Number of female individuals present
	Avg. Number of male individuals present
	Avg. Number of resident families
	Avg. Number of individuals by age group: 0-4 years
	Avg. Number of individuals by age group: 5-9 years
	Avg. Number of individuals by age group: 10-13 years
	Avg. Number of individuals by age group: 14-19 years
	Avg. Number of individuals by age group: 15-19 years
	Avg. Number of individuals by age group: 20-24 years
	Avg. Number of individuals by age group: 25-64 years
	Avg. Number of individuals by age group: > 65 years
	Avg. Number of male individuals
	Avg. Number of individuals male by age group: 0-4 years
	Avg. Number of individuals male by age group: 5-9 years
	Avg. Number of individuals male by age group: 10-13 years
	Avg. Number of individuals male by age group: 14-19 years
	Avg. Number of individuals male by age group: 15-19 years
	Avg. Number of individuals male by age group: 20-24 years
	Avg. Number of individuals male by age group: 25-64 years
	Avg. Number of individuals male by age group: > 65 years
	Avg. Number of female individuals
	Avg. Number of individuals female by age group: 0-4 years
	Avg. Number of individuals female by age group: 5-9 years
	Avg. Number of individuals female by age group: 10-13 years
	Avg. Number of individuals female by age group: 14-19 years
	Avg. Number of individuals female by age group: 15-19 years
	Avg. Number of individuals female by age group: 20-24 years
	Avg. Number of individuals female by age group: 25-64 years
	Avg. Number of individuals female by age group: > 65 years
	Avg. Number of families with 1 unmarried child
	Avg. Number of families with 2 unmarried children's
	Avg. Number of families
	Avg. Number of families with children under 15 years old
	Avg. Number of families with children under 6 years old
	Avg. Number of families with children over 15 years old
	Avg. Number of households with 1 or 2 divisions
	Avg. Number of households with 3 or 4 divisions
	Avg. Number of households between 100m^2 and 200m^2 of area

	Avg. Number of households with more than 200m^2 of area
	Avg. Number of households with less than 50m^2 of area
	Avg. Number of households between 50m^2 and 100m^2 of area
	Avg. Number of rented residences
	Avg. Number of households with water
	Avg. Number of households with bathtubs
	Avg. Number of households with sewage
	Avg. Number of households with a toilet
	Avg. Number of classic family houses with parking for 1 vehicle
	Avg. Number of classic family houses with parking for 2 vehicles
	Avg. Number of classic family houses with parking for 3 vehicles or more

Table 9.2 - List of variables of the "Socio-Economic" group

```
Resampling performance over subset size:

 Variables   RMSE Rsquared    MAE  RMSESD RsquaredSD   MAESD Selected
         2 0.5312   0.7364 0.4150 0.07416    0.07954 0.05393
         4 0.4846   0.8018 0.4072 0.04811    0.05085 0.03522
         6 0.4630   0.8236 0.3780 0.05964    0.06123 0.03902
         8 0.4588   0.8310 0.3796 0.05127    0.05786 0.03605        *
        10 0.4643   0.8277 0.3777 0.05813    0.06896 0.03753
        12 0.4643   0.8285 0.3733 0.05513    0.05983 0.03509
        14 0.4660   0.8290 0.3756 0.05623    0.06411 0.03638
        16 0.4663   0.8294 0.3739 0.05591    0.06416 0.03465
        18 0.4675   0.8298 0.3734 0.05497    0.06620 0.03330
        20 0.4645   0.8324 0.3721 0.05228    0.06141 0.03207
        30 0.4666   0.8330 0.3707 0.05485    0.06242 0.03428
```

Figure 9.2 - RFE output for the "Socio-Economic" group variables

Group	**Variables**
Building Environments	Avg. Number of households
	Avg. Number of collective households
	Avg. Number of collective household's classic
	Avg. Number of collective household's non-classic
	Avg. Number of usual residences
	Avg. Number of vacant dwellings
	Avg. Number of households with 1 or 2 floors
	Avg. Number of households with 3 or 4 floors
	Avg. Number of households with 5 or more floors
	Avg. Number of classic buildings
	Avg. Number of classic buildings with 1 or 2 households
	Avg. Number of classic buildings with 3 or more households
	Avg. Number of buildings in band
	Avg. Number of buildings semi-detached
	Avg. Number of isolated buildings
	Avg. Number of buildings: other
	Avg. Number of buildings built before 1919
	Avg. Number of buildings built between 1919 and 1945

	Avg. Number of buildings built between 1946 and 1960
	Avg. Number of buildings built between 1961 and 1970
	Avg. Number of buildings built between 1971 and 1980
	Avg. Number of buildings built between 1981 and 1990
	Avg. Number of buildings built between 1991 and 1995
	Avg. Number of buildings built between 1996 and 2000
	Avg. Number of buildings built between 2001 and 2005
	Avg. Number of buildings built between 2006 and 2011
	Avg. Number of buildings with stone structure
	Avg. Number of buildings with concrete structure
	Avg. Number of buildings with a sign
	Avg. Number of buildings with structure: other
	Avg. Number of buildings without a sign
	Avg. Number of residential buildings only
	Avg. Number of mainly non-residential buildings
	Avg. Number of mainly residential buildings
	Avg. Number of hospitals
	Avg. Number of health centers
	Avg. Number of pre schools
	Avg. Number of 1st cycle schools
	Avg. Number of 2nd and 3rd cycle schools
	Avg. Number of secondary schools
	Avg. Number of professional schools
	Avg. Number of universities
	Avg. Number of bus stops
	Avg. Number of metro stations
	Avg. Number of train stations
	Avg. Number of piers
	Avg. Number of cultural spots
	Avg. Number of restaurants, coffes, bars and supermarkets
	Avg. Number of bike parks
	Avg. Number of touristic establishments

Table 9.3 - List of variables of the "Building Environments" group

```
Outer resampling method: Cross-Validated (10 fold, repeated 5 times)

Resampling performance over subset size:

 Variables   RMSE Rsquared    MAE  RMSESD RsquaredSD   MAESD Selected
         2 0.6477   0.5983 0.5399 0.11822    0.18332 0.08146
         4 0.5217   0.7545 0.4426 0.05483    0.08950 0.03999
         6 0.4390   0.8289 0.3441 0.06247    0.06820 0.03638
         8 0.4111   0.8591 0.3291 0.04870    0.05147 0.03114
        10 0.4100   0.8600 0.3204 0.05244    0.05642 0.03282
        12 0.4084   0.8626 0.3176 0.04981    0.05122 0.03098
        14 0.4082   0.8633 0.3192 0.04790    0.05023 0.03006        *
        15 0.4111   0.8604 0.3177 0.04981    0.05156 0.03133
        25 0.4104   0.8614 0.3178 0.05048    0.05098 0.03204
```

Figure 9.3 - RFE output for the "Building Environments" group variables

```
Recursive feature selection

Outer resampling method: Cross-Validated (10 fold, repeated 5 times)

Resampling performance over subset size:

 Variables  RMSE Rsquared    MAE  RMSESD RsquaredSD   MAESD Selected
         1 1.099  0.01239 0.8663 0.09494    0.01479 0.08292
         2 1.103  0.03059 0.8771 0.10616    0.03135 0.08020
         3 1.082  0.02829 0.8682 0.09484    0.03520 0.07432
         4 1.062  0.02625 0.8523 0.09035    0.03355 0.06630
         5 1.043  0.02047 0.8366 0.08583    0.02342 0.06241        *
         6 1.056  0.02003 0.8485 0.08994    0.02765 0.06949
```

Figure 9.4 - RFE output for the "Weather conditions and date" group variables

10.ANNEXES

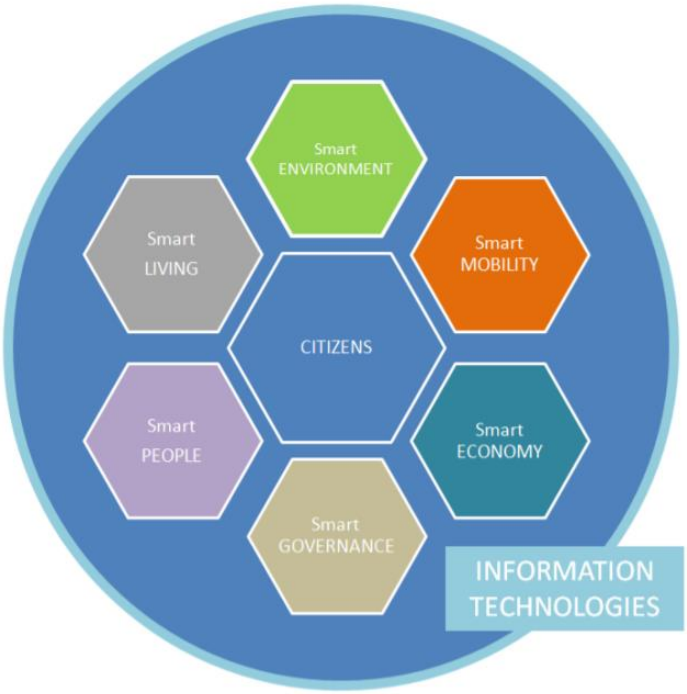


Figure 10.1 - Proposed model for smart cities (Giffinger et al. 2007)

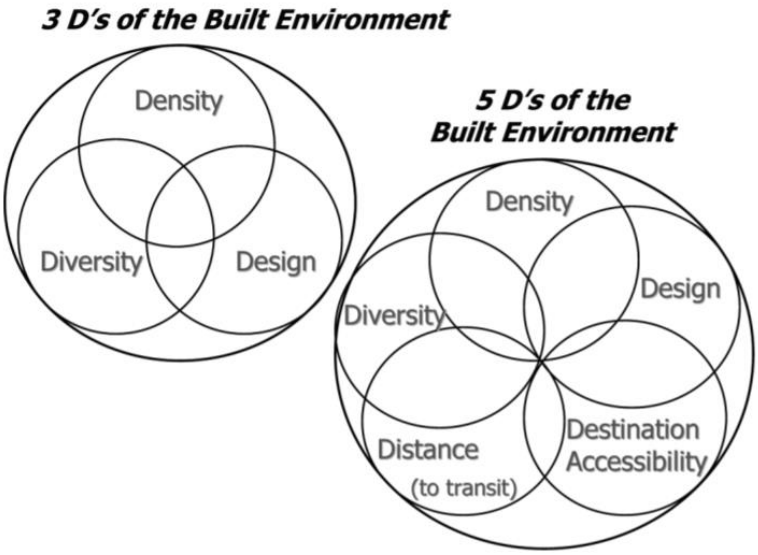


Figure 10.2 - Relationship between built environment and pedestrian mobility - Evolution of the 5 D's (Cervero et al. 2009)