INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Car Accidents: How Much Is Due To External Factors And Conditions? A Data Science Approach For The Portuguese Road Network.

Ana Sofia Gamarro Figo Lourenço

Master In Business Analytics

SUPERVISOR:

Professor Nuno Duarte Fialho Sanches Borges Dos Santos, Guest Assistant Professor, ISCTE Business School, Department of Quantitative Methods For Management And Economics

October, 2022

BUSINESS
SCHOOL

Department of Quantitative Methods for Management and Economics

Car Accidents: How Much Is Due To External Factors And Conditions? A Data Science Approach For The Portuguese Road Network.

Ana Sofia Gamarro Figo Lourenço

Master In Business Analytics

Supervisor:

Professor Nuno Duarte Fialho Sanches Borges Dos Santos, Guest Assistant Professor, ISCTE Business School, Department of Quantitative Methods For Management And Economics

October, 2022

Resumo

O principal objetivo desta pesquisa é encontrar quais e qual o peso dos fatores externos nos acidentes e das vítimas que resultam do mesmo. A variável dos acidentes contém todos os acidentes que aconteceram em Portugal Continental em 2018, de acordo com o INE e a variável das vítimas contém todas as vítimas desde ligeiras, graves e mortais em Portugal Continental em 2018 resultantes de acidentes.

Os dados foram retirados de fontes como a IPMA, PORDATA, INE e Here (rede viária de Portugal). Após a recolha dos dados, a análise dos mesmos e a criação de novas variáveis, com a ajuda dos softwares QGIS e SPSS Statistics, foram todas organizas por município pertencentes ao país em estudo.

Após toda a seleção das variáveis, de acordo com a literatura, foram criados diferentes modelos de forma a retirar conclusões sobre as varáveis (fatores externos). Para este estudo foram criados dois modelos diferentes, para acidentes e vítimas pois estas duas variáveis (targets) não tinham uma forte correlação linear apresentando um valor de Sig de 0,554.

De modo a generalizar para as estruturas rodoviárias portuguesas e para outros países com características semelhantes a Portugal, foi utilizado o método de *bootstrap* como uma estratégia de simulação, deste modo gerou-se 300000 novos dados. Após a avaliação dos dados verificou-se que os fatores externos, utilizados nestes modelos têm uma capacidade explicativa inferior a 50%, mas a dependência espacial é um fator chave e muito importante em problemas geo-espaciais.

Keywords: Acidentes, Tráfego, Vítimas, Dependência Geo- espacial; Índice de Moran

JEL Classification System: R41, C21

Abstract

The main objective of this research is to find out which and what weight external factors have in accidents and victims resulting from them. Within the variable accidents, all accidents that happened in Mainland Portugal in 2018 are counted, according to INE, as the victims include all victims, in Mainland Portugal in 2018 resulting from an accident.

The data used was taken from several sources, namely, PORDATA, IPMA, INE, DGTerritório and Here. After collecting the data, the data was thoroughly analyzed and new variables were created with the help of QGIS and SPSS Statistics software, all of them organized by municipalities belonging to the country under study.

After all the analysis and selection of variables with the Geoda software and the literature, different models were performed in order to draw conclusions about the selected variables. For this study, two different models were made, for accidents and victims (per 1000 meters and per 1000 inhabitants respectively), because these two variables (targets) didn't have a strong linear correlation, presenting a value of 0.036 (Pearson correlation) since there was no relationship between the variables.

In order to generalize to Portuguese road structures and to other countries with similar characteristics to Portugal, the bootstrap method was used as a simulation strategy, thus generating 300,000 new data. After evaluation the data, it was found that the external factors used in these models have an explanatory capacity of less than 50%, but spatial dependence is a key and very important factor in geospatial problems.

Keywords: Car accidents, traffic, victims, spatial dependence, Moran's Index

JEL Classification System: R41, C21

Index

Table Index

Figure Index

Chapter 1 – Introduction

This investigation will focus on the study and identification of the external factors that may influence the number of car accidents in Mainland Portugal. The main objectives of this study are identifying the most relevant factors and quantify how relevant the external factors are.

In several countries there are many studies about road accidents and how to predict them, but most of the articles refer as the main problem the driver or the vehicle conditions, but, on the other hand, it is also mentioned that the traffic on the road and the geometry of the road are factors that contribute to accidents (Xu et al., 2020), as well as the poor lighting on the roads (Shweta et al., 2021) and how the road is presented - dry or wet (Chong et al., 2004).

According to Ameen and Naji (2000) it is important to identify the causes of road accident fatalities because the growth of technology, population and consequently the number of vehicles and their use, with that, it's possible that more accidents might happen created by traffic, so it is important to solve accidents as soon as possible so that others do not happen consequently (Dogru & Subasi, 2012). In addition, many authors study the causes of road accidents but it's difficult to get a universal model because of the environment and geographical changes within different regions (Ameen & Naji, 2000).

According to the World Health Organization, recent assessments show that traffic accidents are responsible for more than one million deaths per year and are the largest public health problem and socio-economic cost according to Albuquerque, et al. (2021) and Shweta, et al. (2021).

Additionally, the accident rate in Mainland Portugal had a 4.6% growth between 2008-2018, meaning an increase of 36162 accidents in 2018, which resulted in 46034 victims (lightly injured, seriously injured and fatalities), 704 fatalities (Marktest - Sales Index/INE, 2021) which represents a weight of 1%, that might seem a low percentage, but it is alarming considering that many of these accidents could probably be avoided and thus preventing unwanted deaths.

Accidents on the roads are both an economic and a social problem. As we all know, in today's world, any event involves monetary issues and accidents are no exception. Speaking more specifically of insurance in the first and second quarter of the year 2021, 488,824 million euros were spent on automobile claims and in 2020 a total of 995,783 million euros, according to the Insurance Supervision Authority and Pension Fund (ASF, 2022). In addition, we have the social issue, any type of accident can bring a nuisance to people especially the injured resulting from it.

These accident figures can be related to several reasons: the driver, the vehicle itself and/or external factors such as the state of the road, the weather, the geography of the road among others. As we have seen above, these factors can be determinant of both the type and severity of accidents. With a review of the studies conducted in Portugal it was found that this focus more on human factors, since it is the most determinant (Pereira, 2016), or in the creation of a profile of the type of driver with more accidents (Bon de Sousa et al., 2016). But on the other hand, some studies take a more micro view in relation to some external factors, such as Guerreiro (2008) who conducted a descriptive study on the accident rate in Portugal and the reason for accidents on the EN6 and A5, suggesting some corrective measures, and Ilharcos, et al. (2013) who studied intersections, roundabouts and segments in the city of Lisbon, referring that there was a great difficulty in obtaining data and that the methodology used was not the most correct, this being initiated by explanatory variables of traffic and characteristics of the geometry of the road organized by 3 types of elements: roundabouts, segments and intersections with 3 or 4 lanes.

All the work done in this area uses only one or two external factors, mainly road geometry or road brightness, and the studies are restricted to a small area and not to the entire road network of the country under study.

As mentioned above, the accident rate between 2008 and 2018 has a growth rate, that is, an increase in accidents on Portuguese roads that could result in more injuries and deaths that many of them, with the right measures, could be avoided.

Chapter 2 – Literature Review

2.1) Protocol for the systematic literature review

In order to consolidate the literature review, a protocol was followed to understand how external factors can influence traffic accidents. For this, a main question and two specific questions were established. The main question is it is possible to explain accidents based on external factors; and the specific questions are: do external factors influence accidents/ victims and how much do external factors influence accidents/victims.

In addition, it will also analyze which external factors are the most important and how the available articles come to these conclusions. Thus, articles referring to accidents and accident victims were removed from B-on (https://www.b-on.pt/). All these articles are within a 22-year period, between 2000 and 2022, and must meet several inclusion and exclusion criteria, such as:

Inclusion criteria:

- Articles published in academic journals;
- Business and Economics articles.

Exclusion criteria:

- Articles that are not complete;
- Literature reviews;
- Duplicate articles;
- Engineering articles.

The search is performed using all of the above criteria and a query formed with different keywords for this research.

The query formed for this research is: ("car accident" OR "car disaster" OR "automobile accident" OR "auto accident" OR "road accident") AND ("external factor*") AND ("machine learning" OR model* OR predictive* OR segmentation* OR ml OR analytics OR forecasting*).

2.1.1) Selection Procedures

The articles selected were based on the above criteria: query, language, period, and inclusion and exclusion criteria. After the selection of these criteria, they were evaluated from a reading of the abstract to verify if they correspond to the objectives set for this systematic literature review,

relating traffic accidents to external factors. The scientific articles were then read and evaluated according to the criteria described in the following section: Selection Criteria.

The articles were evaluated according to the quality evaluation questions, these have answers as Yes, No and Partially.

2.1.2) Selection Criteria

ID	Quality Criterion	Possible Answer
Q1	Does it explain the importance of the study?	Yes/No/Partially
Q2	Does it address the importance of external factors?	Yes/No/Partially
Q3	Does it compare the different methods?	Yes/No/Partially
Q4	Describe the different methods used?	Yes/No/Partially
Q5	Does it have a good methodology?	Yes/No/Partially
Q6	Does it use external factors as a variable?	Yes/No/Partially
Q7	Does it describe the data processing?	Yes/No/Partially
Q8	What is the purpose of the model used?	Yes/No/Partially
Q9	Does it describe the evaluation steps of the model?	Yes/No/Partially

Table 2.1 - Quality criteria for evaluating the articles under study

2.1.3) Articles integrated into the systematic literature review

The table below shows a list of the articles selected for the systematic literature review and some that were found externally and relevant to the study.

ID	YEAR	TITLE	AUTHORS	JOURNAL
1	2000	Causal models for road accident fatalities in Yemen	Ameen, J. R. M. & Naji, J. A.	Accident Analysis and Prevention
2	2017	A hybrid clustering and classification approach for predicting crash injury severity on rural roads.	Hasheminejad, S. H., Zahedi, M. & Hasheminejad, S. M. H.	International Journal of Injury Control and Safety Promotion
3	2020	A traffic prediction model based on multiple factors	Wang, J. & Chen, Q.	The Journal of Supercomputing
4	2020	Machine learning models and techniques for VANET based traffic management: Implementation issues and challenges.	Khatri, S., Vachhani, H., Shah, S., Bhatia, J. Chaturvedi, M., Tanwar, S. & Kumar, N	Peer-to-Peer Networking and Applications
5	2020	Traffic Prediction Using Multifaceted Techniques: A Survey.	George, S. & Santra, A. K.	Wireless Personal Communications
6	2020	Why Is Artificial Intelligence Blamed More? Analysis of Faulting Artificial Intelligence for Self-Driving Car Accidents in Experimental Settings	Hong, J. W.	International Journal of Human-Computer Interaction

7	2021	Smart Cities: Data-Driven Solutions to Understand Disruptive Problems in Transportation.	Albuquerque, V., Oliveira, A., Barbosa, J. L., Rodrigues, R. S., Andrade, F., Dias, M. S. & Ferreira, J.C.	Energies
8	2022	A hybrid neural network for driving behavior risk prediction based on distracted driving behavior data	Fu, X., Meng, H., Wang, X., Yang, H. & Wang, J.	PLOS ONE

Table 2.2 - Articles selected for the systematic review of the literature

ID	YEAR	TITLE	AUTHORS	JOURNAL
9	2005	Traffic Accident Analysis Using Machine Learning Paradigms	Abraham, A. & Paprzycki, M.	Informatica
10	2010	Effect Of Vehicle Characteristics On Crash Severity: Portuguese Experience	Torrão, G., Coelho, M. & Rouphail, N.	WCTR
11	2012	Traffic Accident Detection By Using Machine Learning Methods	Dogru, N. &Subasi, A.	Information Systems and Sustainability
12	2021	A Framework for Analysing Road Accidentes Using Machine Learning Paradigms	Shweta, Yadav. J, Batra, K. & Goel, K.	Journal of Physics: Conference Series
13	2021	Predictive Modeling of Maximum Injury and Potencial Economic Cost in a Car Accident Based on the General Estimates System Data	Alkan, G., Farrow, R., Liu, H., Moore, C., Keung, H., Ng, T., Stokes, L. Xu, Y., Xu, Z. Yan, Y & Zhong, Y.	Computational Statistics
14	2020	Analysis of the Risk Dactors Affecting the Severity of Traffic Accidents on Spanish Crosstown Roasd: The Driver's Perspective	Casado-Sanz, N., Guirao, B. & Attard	Sustainability

Table 2.3 - Important articles for the literature review, external

2.2) Car Accidents and Victims

According to the World Health Organization (2022), recent assessments show that traffic accidents are responsible for more than one million deaths per year and are the largest public health problem and socio-economic cost according to Albuquerque, et al. (2021) and Shweta, et al. (2021). Road accidents not only include material loss, injuries, and deaths, they can have high costs to governments, including economic, social and political (Hasheminejad et al., 2017).

Thus, identifying the causes of accidents is quite important especially with the growth in the number of vehicles, population, technology and other factors (Ameen & Naji, 2000). Driver behavior is known to have a great influence and significance in traffic accidents according to Fu, et al. (2022) and Torrão, et al. (2010).

The identification of accidents and their causes is very important because by identifying some of their causes and their weight we can reduce accidents, and identifying them will reduce traffic and driver delay, improve road safety and especially avoid other accidents by the large amount of

traffic created by an accident, so it is necessary to solve accidents as soon as possible so that others do not happen consequently (Dogru & Subasi, 2012).

Many authors try to generalize the causes of road accidents, but the problems have different trends because each country/city presents its own geography, environment and roadway (Ameen & Naji, 2000).

Although there are the mentioned differences in all the articles under study, one serious problem should be highlighted: traffic monitoring and congestion on the roads. High traffic on roads translates into serious problems of congestion, safety, environmental impact (Wang & Chen, 2020) and an increase in the number of incidents (Khatri et al., 2020). Thus, "reducing traffic accidents is a crucial social problem" (Wang & Chen, 2020).

It should also be noted that there are several factors that directly affect road accidents, mainly 4 factors: human, roadway, environmental and the vehicle (Hasheminejad et al., 2017).

"In 2008, 50% of the world's population lived in urban areas, and it was growing exponentially. By 2050, 70% of the world's population is expected to live in metropolitan areas" (Albuquerque et al., 2021). This transition is quite notorious in Portugal, more and more of the Portuguese population is moving to urban areas / large cities of the country, leaving the interior. This migration, to big cities increased pollution, more traffic on the roads and, in turn, road accidents in large cities (Albuquerque et al., 2021).

Additionally, car accidents can result in a major tragedy, namely, fatalities, and can also result in minor or serious injuries. Statistically, the highest number of fatalities happen in urban areas and with pedestrians, because they have no protection when there is a collision with a vehicle, which can result in a more serious outcome (Casado-Sanz et al., 2020).

2.3) External Factors

Accidents happen based of several factors that can influence the existence or non-existence of accidents, or even their severity. The main factors are internal factors such as the driver. Then we have external factors such as road conditions, weather conditions, road geography, among others (Hong, 2020).

It is known that the driver is often indicated as the main responsible in accidents, and to a lesser extent, the external factors. After a simulation study using 284 participants comparing accidents with drivers or with artificial intelligence systems, drivers place greater blame on

external factors when accidents happen with cars driven by artificial intelligence than with drivers (people) (Hong, 2020).

According to Shweta, et al. (2021) it is essential to understand the data regarding accidents, to detect the burden of damage and the source of the problem, namely, the roadway, to be able to provide the necessary safety to drivers and databases regarding this data are essential to demystify this problem. Furthermore, it is concluded that low light or dark lighting contributes to the causes of accidents compared to other regions, in this case in Canada (Shweta et al., 2021).

Although external factors have less weight, they always have some and by finding out which factors are more influential one can avoid some (Hong, 2020).

For this thesis, external factors such as road conditions, weather, lighting, low lighting according to Albuquerque, et al. (2021) and Chong, et al. (2005) and time of day are examples of factors that can make a big difference in traffic accidents (Fu et al., 2022) and the injury severity (Shweta et al., 2021). Furthermore, it is concluded that low light or dark lighting contributes to the causes of accidents compared to other regions, in this case in Canada (Shweta et al., 2021).

In addition to the factors described above, there are other variables external to the drivers that one needs to pay attention to, such as the social and economic diversity that exists in the country under study, which varies from country to country (Ameen & Naji, 2000). Historical problems (as happened in Yemen between 1989 and 1990) can also affect the number of accidents and their severity (Ameen & Naji, 2000), so it is necessary to pay special attention to outside influences, such as regional instability and internal politics, as in the case of Yemen (Ameen & Naji, 2000).

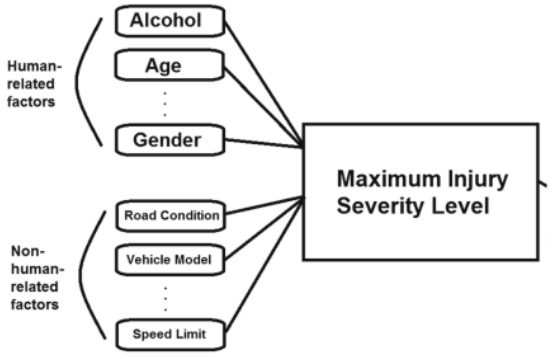


Figure 2.1 - Factors Related to human being and external factors to severity

Source: Alkan, et. al. (2021)

2.4) Models used in accident prevention and victims

Various algorithms have already been used for the study of traffic accident prevention such as: predictive models, statistical models, decision making systems, accident severity prediction using data mining, artificial neural networks, and support vector machine to determine which key factors affect accident severity according to Hasheminejad, et al. (2017) and Wang and Chen (2020).

According to Fu, et al., (2022) it is possible to divide the models used for the study of traffic accident prediction into 3 categories:

- Models based on time series and Kalman filtering models;
- Nonlinear statistical models, based on non-parametric regression and chaos theory;
- Models based on machine learning, specifically neural networks and support vector machine (SVM).

In order to relate fatalities to road accidents, a model called the Smeed Model was created. This model has as its main objective to relate mortality rates and the number of vehicles per 10 km by population size (Ameen & Naji, 2000).

Source: Ameen and Naji (2000)

$$\frac{F}{V} = a\left(\frac{V}{P}\right)^{-b} \qquad (1)$$

The above equation includes several components at the monitoring level and income level of developing countries. Thus, F is the number of fatalities, V is the number of vehicles on the road per 10km and P is the population size, with a and b being constants of the equation (Ameen & Naji, 2000).

Because of the way the Smeed model is implemented there are several authors criticizing it, indicating it is a model only for developing countries, such as the case of the study (Yemen)(Ameen & Naji, 2000).

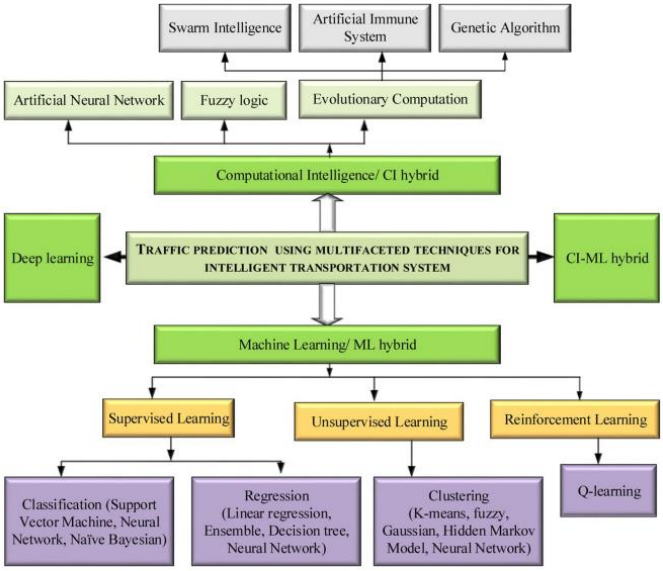


Figure 2.2 - Traffic Prediction using multifaceted Technique

Source: George and Santra (2020)

Each technique, like artificial neural network, fuzzy logic, SVM, neural network, linear regression, decision tree, k-mean, was placed within 4 categories: Machine Learning (ML), Computational Intelligence (CI), Deep Learning (DL) and hybrid algorithms (George & Santra, 2020).

Thus, the categories where external factors have already been considered, using historical data and real-time data, are Machine Learning, Computational Intelligence and in the Computational Intelligence Hybrid algorithms (George & Santra, 2020).

It should be noted that models that use, especially, neural networks have had great advances and great results in accident prediction, but it is necessary to pay attention that most of the studies only consider the vehicles or drivers, others only study the temporal space of accidents and do not combine it with the spatial perspective and, finally, most of the studies obtain data from simulations and not real data (Fu et al., 2022).

According to Torrão, et al. (2010), C&RT is a model in which you can choose the independent and dependent variables that can give a great explanatory power of accidents and victims. It is a model that does not need a predefined relationship between the independent and dependent variables, being classified as an advanced data mining technique (Torrão et al., 2010).

Shweta et. al. (2021) divided the database by analyzing in clusters, using the k-means cluster technique with 4 clusters, the categories used were, district (location), lighting, visibility, and road

conditions, identifying that aggressive and distracting driving was one of the main causes for the increase in accidents, totaling 62.9%. This cause is also included as one of the main causes of accidents along with pedestrians.

Chapter 3 – Methodology and Data Sources

According to Razein, et al. (2016), many road accidents studies use the Cross Industry Process for Data Mining (CRISP-DM) methodology, a powerful methodology for data mining.

"The CRISP-DM (Cross Industry Standard Process for Data Mining) project addressed parts of these problems by defining a process model which provides a framework for carrying out data mining projects which is independent of both the industry sector and the technology used. The CRISP-DM process model aims to make large data mining projects, less costly, more reliable, mor repeatable, more manageable, and faster." (Wirth & Hipp, 2000).

The chosen methodology contains 6 consecutive phases, although it is not mandatory to follow all phases rigidly, CRISP-DM is a very complete methodology that pays full attention to all the necessary parts to perform a good data mining study, starting with the business study, and ending with the implementation (Chapman et al., 2000).

To carry out this study, we followed the approach illustrated in the figure 3.1, following the CRISP-DM methodology.

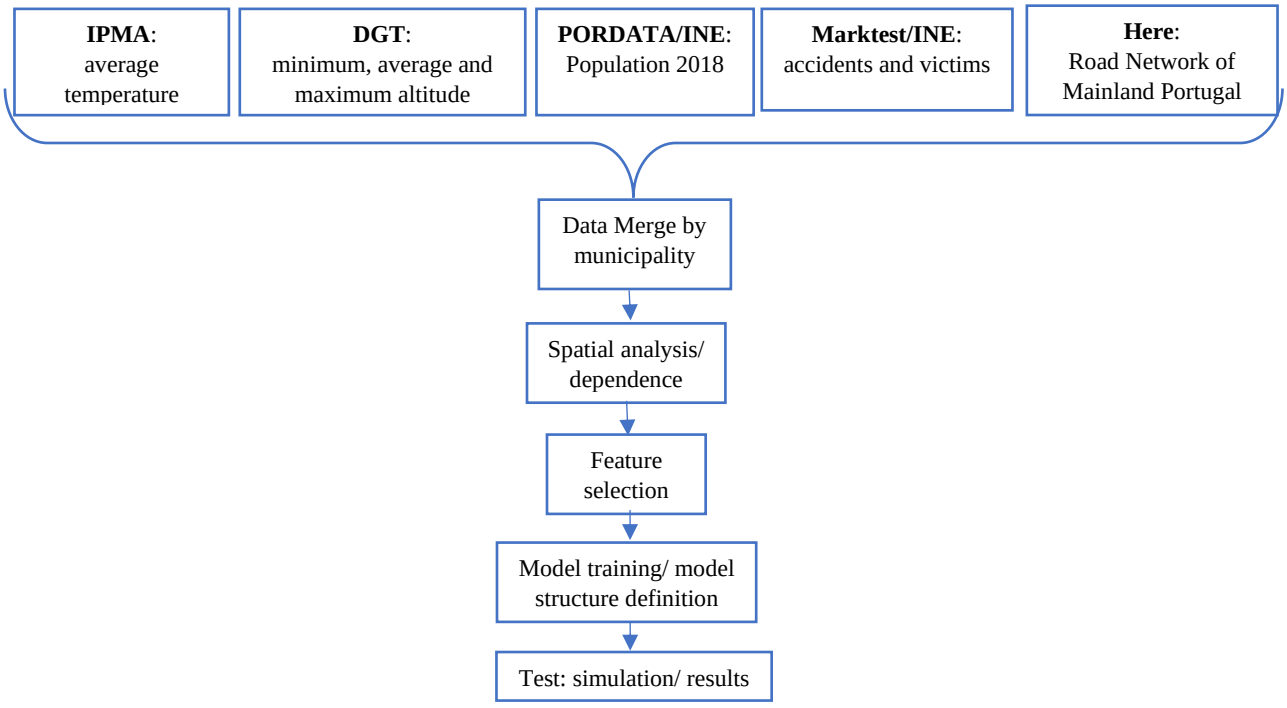


Figure 3.1 – Approach used: flowchart

In order to support this methodology, many software programs were used for spatial data analysis, like QGIS, GeoDa and GeoDa Space. In addition, to create new variables, to do all the

modifications and all of models, both SPSS Statistics and SPSS Modeler were used. These two software's were developed by IBM and adapted to the CRISP-DM methodology with several successes in solving problems in several companies in different industries (Chapman et al., 2000).

The data used for this study were collected from a variety of reliable sources.

IPMA is an institution belonging to the portuguese state from which data such as the thermal amplitude of the municipalities were taken (IPMA, 2022)

PORDATA is a database organized and developed by the Francisco Manuel dos Santos Foundation that prioritizes the "collection, organization, systematization and dissemination of information on multiple areas of society, for Portugal, municipalities and European countries. The statistics disclosed are from official and certified sources". (PORDATA, 2022). In this source there are several studies and one of them is the existing population in mainland Portugal of 2018 divided by municipalities, which was taken for this study.

The Direção Geral do Território (DGT) is a portuguese company owned by the portuguese state. The portuguese territory as the minimum, maximum, and average altitude of each municipality were taken from this institution for this study. (DGT - Direção Geral do Território, 2022)

MARKTEST is a group consisting of several companies specializing in market research. One of its market studies is the Sales Index, from which the data on road accidents and victims for 2018 was taken. (MARKTEST, 2022).

INE is a statistical studies company, independently and impartially. (INE, 2022)

Finally, Here is a location data and technology company that creates digital maps and has a strong presence in the automotive industry (Here, 2022). Through Here it was possible to obtain a database with several characteristics about the road segments present in Mainland Portugal.

Chapter 4 – Statistical Analysis of Data

4.1) Statistical Analysis of Data

The data collected from PORDATA/INE were organized by the 278 municipalities belonging to Portugal's Mainland.

The Here network was divided into links, that is, each road segment, sidewalk, or road existing in continental Portugal. Thus, the Here network totaled approximately 2 million links, of which 124,127 belonged to the North Road network and 108,800 to the South Road network, as can be seen in the table and in the image below.

Statistics		
Road Network of Mainland Portugal		
N	Valid	2324127
	Missing	0

Table 4.1 - Road Network Data

Source: Here

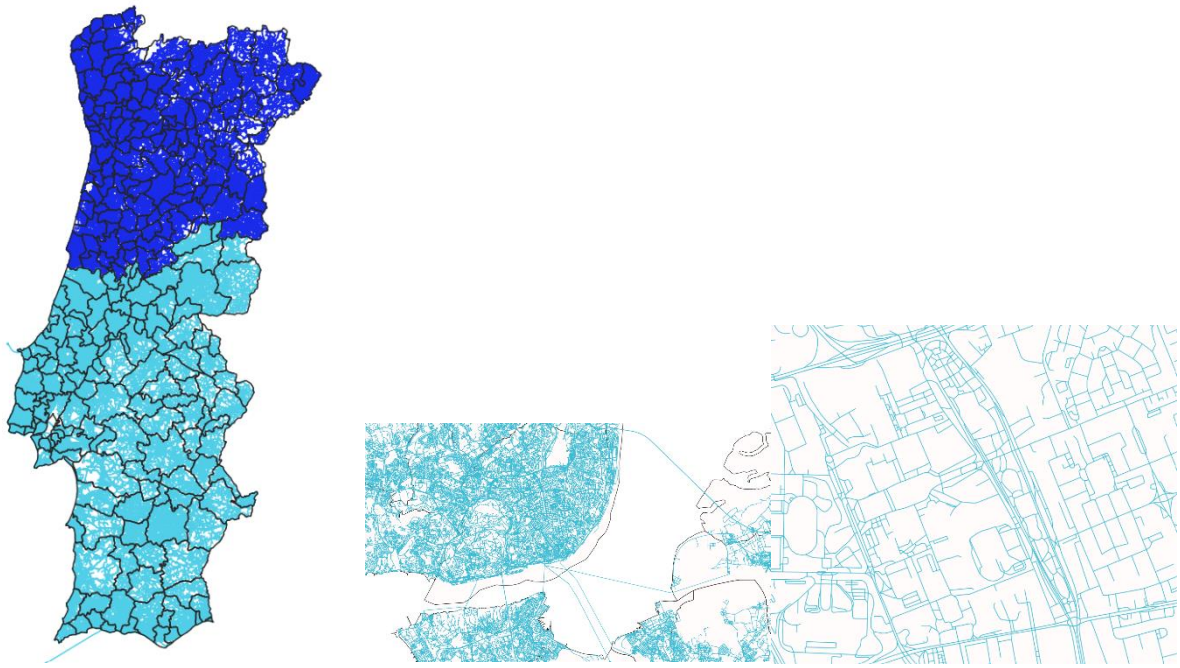


Figure 4.1 - Road Network Distribution in Portugal , University city and Campo Grande

Source: Here

The Portuguese Road network database contains 108 explanatory variables, like if cars, boats, motorcycles, pedestrians and other are allowed in the road segment or the number of lanes in the

road segment, street name, municipality, district, intersections and others, describing each road link/segment.

Finally, from PORDATA/INE, DGT and IPMA was taken data for the population of 2018, the minimum altitude, average altitude, maximum altitude, and average temperature range for each municipality.

After collecting all the data, we obtained 5 different databases. All of them had to be changed in order to later unite all variables in the same database, organized by municipalities. In the database of the road network of mainland Portugal there were several variables that were not present in the known literature. Subsequently, each variable had its own way of reading, some were continuous, others nominal and others numerical.

Given that we are dealing with spatial data, most variables are sensitive to spatial units size and importance. For instance, one cannot compare the number of accidents in a small spatial unit and in a large one. Thus, new relative variables had to be created so that the original variables were comparable and conclusions could be drawn. We started by creating a variable called length, that is, we measured (in meters) the length of each link and in this way, we aggregated the entire database by municipality and join all of the data bases. Later, the variables were altered to put the same way of reading, that is, with the same denominator (divided by the length variable), only the variable referring to victims was put as divisor the population, because the victims are related to the population and not with the length of road of each municipality.

The dependent variables are constituted by the variables related to accidents and victims, as shown in the table below, annex B.

After the selection of variables (based on the literature) and the creation of new variables, we ended up with 67 variables, as shown in the Table 4.2. Still, it was necessary to perform a correlation matrix in order to verify the relationship between the variables and how to group them in order to reduce the number of variables without losing information. The correlation matrix can be found in the appendix (annex C).

After the correlation analysis between independent variables, it was found which variables are linearly related (annex C) and simplify and reduce the number of variables.

Regarding the independent variables it was found that the variables tot_ac_p_comp and tot_vit_p_pop encompassed all other variables of accidents and victims.

	N	Mean	Std. Deviation
alt_min	278	85.0288	116.32886
altitude_med	278	268.6897	213.86185
alt_max	278	648.2158	411.34681
amplitude_media_temperatura	278	22.5831	3.18798
vel6_p_comp	278	287.1519	160.94112
vel7e8_urban_p_comp	278	91.6023	64.32470
vel2e3_urban_p_comp	278	1.7082	4.89777
vel4e5_urban_p_comp	278	5.9338	6.11842
vel6_tunel_p_comp	278	.0318	.25828
intersecoes_p_comp	278	1.1528	1.16323
fun1_p_comp	278	12.3900	24.52306
vel4e5_tunel_p_comp	278	.0327	.37481
fun4_p_comp	278	104.2378	40.15510
auto_moto_p_comp	278	1966.7613	49.87992
vel2e3_tunel_p_comp	278	.0887	.52947
fun2_p_comp	278	20.7701	24.28690
fun3_p_comp	278	48.1596	28.86601
vel7e8_tunel_p_comp	278	.0029	.02819
ped_rodo_p_comp	278	1016.8972	267.70341
Poupulação2018	278	35179.2302	57227.79425

Table 4.2 - Variables choose by correlation matrix

Source: Here, PORDATA/INE, DGT and IPMA

4.1.1) Population

The Population of mainland Portugal by municipalities shows a minimum of 1645 citizens, in 2018 in a single municipality (Barrancos) and a maximum of 507220 citizens, in 2018 (Lisbon). The municipalities with the largest population are concentrated near the large cities, such as Porto, Lisbon and Faro and the surrounding municipalities. This is one of the consequences of the migration of citizens to the big cities, as the interior of Portugal has the lowest number of citizens per municipality (Figure 4.2). There is an average of 35179.23 citizens per municipality, taking into account that 50% of the municipalities have a value below the average, 14626 citizens per municipality, which is represented by their median (Table 4.3).

In addition, 25% of the municipalities have below 6779 citizens and 25% have figures above 38404 citizens per municipality (Table 4.3).

Statistics		
População2018		
N	Valid	278
	Missing	0
Mean		35179.2302
Median		14626.0000
Std. Deviation		57227.79425
Minimum		1645.00
Maximum		507220.00
Percentiles	25	6779.0000
	50	14626.0000
	75	38404.2500

Table 4.3 - Analysis of População2018 variable

Source: INE

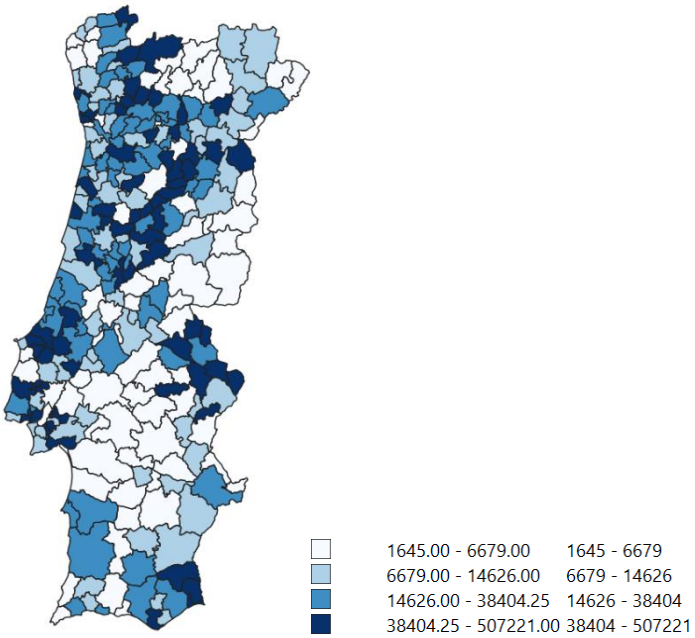


Figure 4.2 - Analysis of the variable população2018

Source: INE

4.1.2) Altitude

The altitude of the municipalities is shown in meters. Thus, within this large group we have 3 variables: the minimum altitude of each municipality, the average altitude of each municipality, and the maximum altitude of each municipality.

As there are no missing values, we can say that we have all the values of altitudes assigned to each municipality.

Statistics				
		alt_min	altitude_med	alt_max
N	Valid	278	278	278
	Missing	0	0	0

Table 4.4 - Valid number of alt_min, altitude_med and alt_max

Source: DGT

Minimum Altitude

The minimum altitude of the municipalities is 43 meters below the average sea level, and the maximum that a municipality can accept as a minimum altitude is 524 meters above the sea level (Table 4.5).

Additionally, looking at the figure for mainland Portugal, it should be noted that the municipalities further west and near the coast have a lower minimum altitude when compared to the interior of mainland Portugal. Furthermore, the highest minimum altitudes are found in the northeastern part of mainland Portugal.

From the percentiles, it was found that 25% of the municipalities are below mean water level, at least 1 meter below. The municipalities with the highest minimum altitude range between 134 and 524 meters above mean sea level (Figure 4.3).

Statistics		
alt_min		
N	Valid	278
	Missing	0
Mean		85.0288
Median		40.0000
Std. Deviation		116.32886
Variance		13532.404
Minimum		-43.00
Maximum		524.00
Percentiles	25	-1.0000
	50	40.0000
	75	134.0000

Table 4.5 - Variable alt_min

Source: DGT

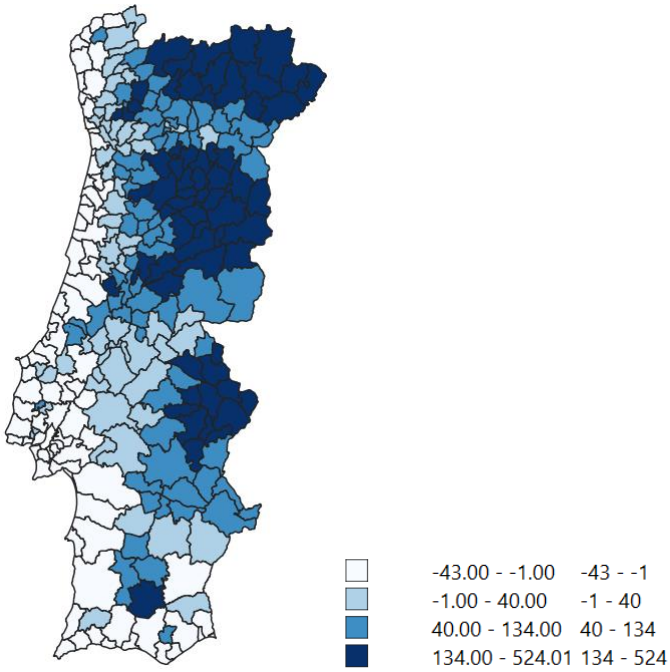


Figure 4.3 -Distribution of alt_min

Source: DGT

Average Altitude

The average altitude of the municipalities varies between 6.95 meters and 827.77 meters above sea level, with an average of 268.69 meters.

It is noteworthy that the areas with the lowest average altitude are along the coast and the most rugged areas are in the interior of mainland Portugal, especially the interior north, where the red color is visible (higher average altitude values).

All municipalities are above the sea level, while 25% of the municipalities have an average altitude below 75.9 meters, 25% have an average altitude above 423.9 meters. The median average altitude is 216.8 meters.

Statistics		
altitude_med		
N	Valid	278
	Missing	0
Mean		268.6897
Median		216.7653
Std. Deviation		213.86185
Variance		45736.893
Minimum		6.95
Maximum		827.77
Percentiles	25	75.8802
	50	216.7653
	75	423.8642

Table 4.6 -Analysis of the altitude_med variable

Source: DGT

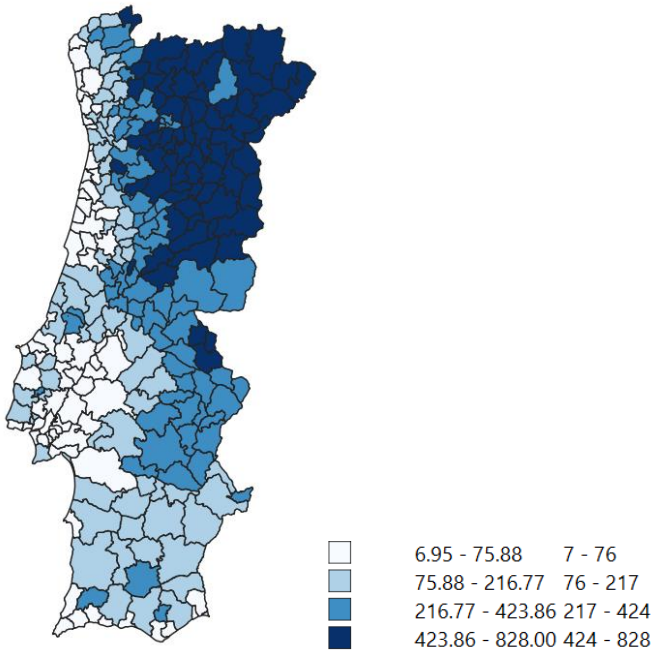


Figure 4.4 - Distribution of the variable altitude_med

Source: DGT

Maximum Altitude

The highest maximum altitudes are found in the center interior and north interior of mainland Portugal and the lowest maximum altitudes along the coast and further south in Continental Portugal with the colors blue and green. This variable presents a high value range of 1958.00 meters, with a minimum value of 35 meters and a maximum value of 1993 meters.

Additionally, the average values of the maximum altitude variable are 648.21 meters, with no missing values.

Statistics		
alt_max		
N	Valid	278
	Missing	0
Mean		648.2158
Median		537.5000
Std. Deviation		411.34681
Variance		169206.199
Minimum		35.00
Maximum		1993.00
Percentiles	25	318.7500
	50	537.5000
	75	957.5000

Table 4.7 - Analysis of the alt_max variable

Source: DGT

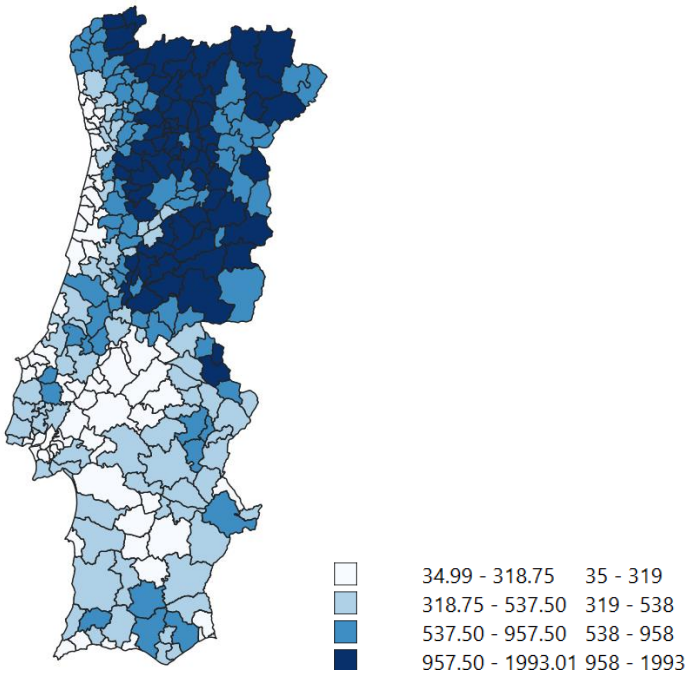


Figure 4.5 - Distribution of the variable alt_max

Source: DGT

4.1.3) Average Temperature Range

The average temperature range was derived as the average of the difference between the highest recorded temperature and the lowest recorded temperature in each municipality. Thus, it is noted that the highest and lowest temperatures are closer together along the coastal coast, mainly in the district of Lisbon and Leiria.

The highest average temperature range is 28,84 ºC, and the lowest is 13,17 ºC, totaling a maximum difference of 15,67 ºC.

The average of the average temperature range corresponds to 22,58 ºC, very close to the median, 22,99 ºC.

With the help of the percentiles, we see that the average temperature range is not very variable, since 25% of the municipalities have a maximum range of 20 ºC, 50% of the municipalities have a maximum range of 23 ºC, approximately. Only 25% of the municipalities have an average temperature range between 25 and 29 ºC.

Statistics		
amplitude_media_temperatura		
N	Valid	278
	Missing	0
Mean		22.5831
Median		22.9880
Std. Deviation		3.18798
Variance		10.163
Minimum		13.17
Maximum		28.84
Percentiles	25	20.1543
	50	22.9880
	75	25.1320

Table 4.8 - Analysis of the variable amplitude_media_temperatura

Source: IPMA

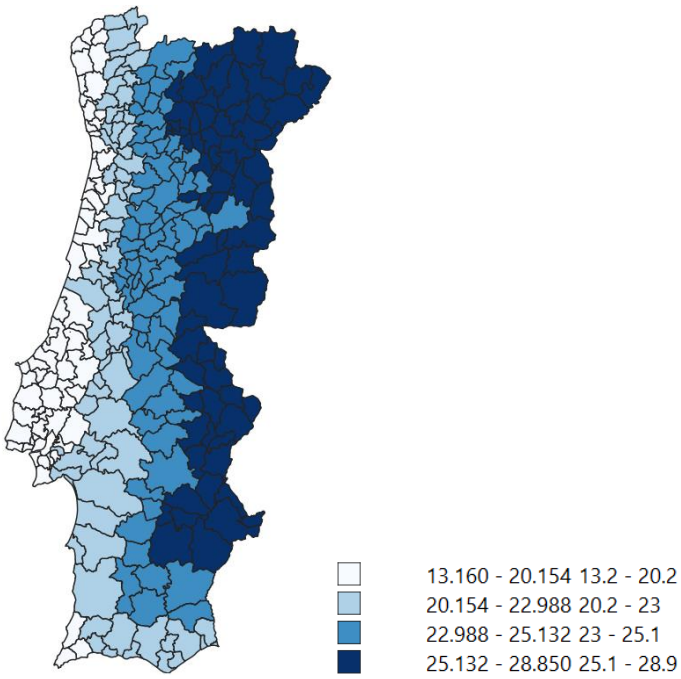


Figure 4.6 - Distribution of the variable amplitude_media_temperature

Source: IPMA

4.1.4) Functional Class

Functional Class serves to rank roads depending on the speed, importance and connectivity of the road and is presented in values between 1 and 5.

Values between 1 and 5 are classified as (Here, https://www.here.com/, viewed on 15/08/2022):

- 1: allows a large volume of traffic movement at maximum speed;
- 2: allows a high volume of traffic movement at high speed;
- 3: allows a high volume of traffic movement;
- 4: allows a high volume of traffic movement at moderate speed between neighborhoods;
- 5: segments whose volume and traffic movement is lower than the other functional classes.

So, we created a new variable that corresponds to how many segments of each functional class there are per 1000 meter (thousand meters). Thus, we verify that there are many more segments of functional class 5, reaching an average of 778,64 per 1000 segments.

The segments with functional class 1 are those where there is a large volume of traffic at maximum speed. With the statistical analysis, analysis of Table 4.9 and Figure 4.7, it can be seen that there are not many segments of functional class 1, since at least 50% of the municipalities do

not contain links of functional class 1. In addition, the municipalities that contain the highest number of segments of per thousand meters (100.00 - 150.00) are in the district of Lisbon.

Furthermore, only one county contains more than 150.01 segments per thousand meters, having 204.25 functional class 1 segments per thousand meters.

Statistics		
fun1_p_comp		
N	Valid	278
	Missing	0
Mean		12.3900
Median		.0000
Std. Deviation		24.52306
Minimum		.00
Maximum		204.25
Percentiles	25	.0000
	50	.0000
	75	16.7542

Table 4.9 - Analysis of variable fun1_p_comp

Source: Here

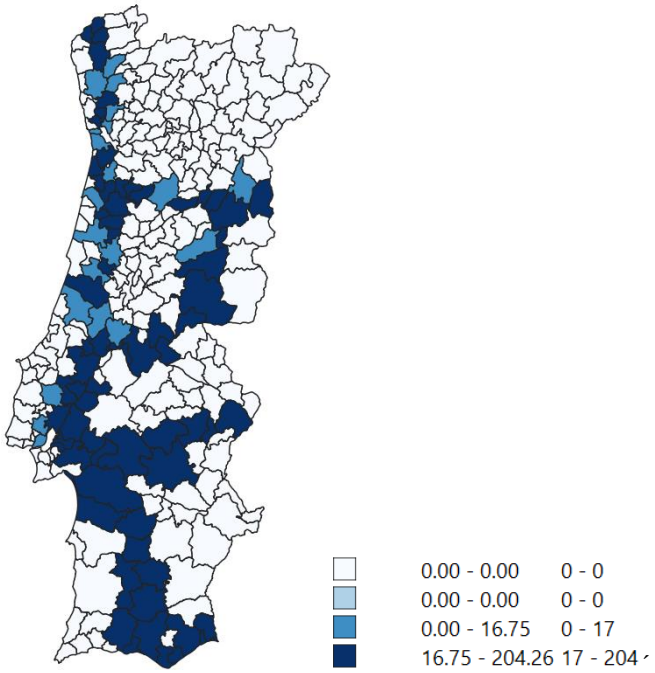


Figure 4.7 - Distribution of variable fun1_p_comp

Source: Here

The segments with functional class 2 represent the segments with high traffic volume and with a high speed. Compared to the number of functional class 1 segments, it can be seen that there is a growth in the number of segments in municipalities with these characteristics, functional class 2, since only 25% of the municipalities do not have segments with this characteristic.

On average municipalities have 20.77 functional class 2 segments per 1000 meters, with a maximum of 278.54 functional class 2 segments per 1000 meters. Furthermore, 50 % of the municipalities have more than 45.43 functional class 2 segments per 1000 meters. The municipalities presenting the highest values of segments of this variable are found mainly in the north of the country with values between 65.46 and 278.5 functional class 2 segments per thousand meters.

Statistics		
fun2_p_comp		
N	Valid	278
	Missing	0
Mean		20.7701
Median		13.5628
Std. Deviation		24.28690
Minimum		.00
Maximum		105.27
Percentiles	25	.0000
	50	13.5628
	75	34.1132

Table 4.10 - Analysis of variable fun2_p_comp

Source: Here

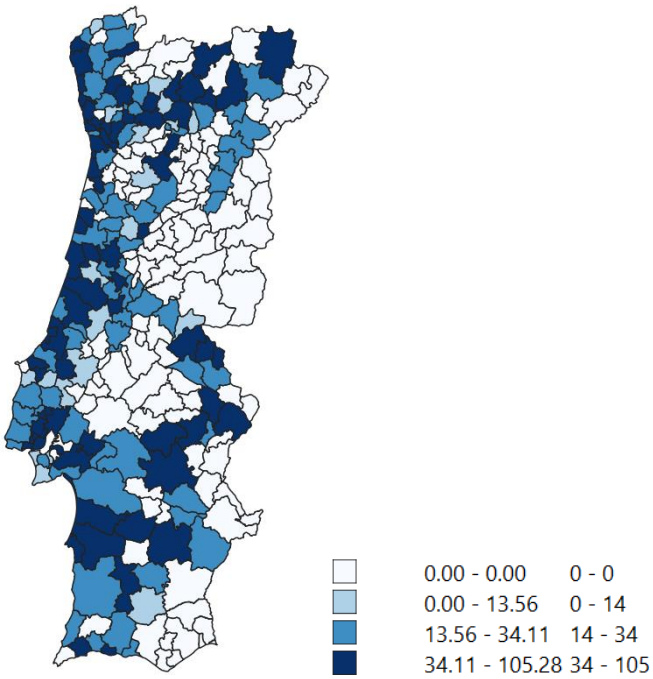


Figure 4.8 - Distribution of variable fun2_p_comp

Source: Here

Functional class 3 represents segments with a large volume of traffic. As it was verified before, functional class 3 segments are in larger number than functional classes 2 and 3. This can be verified since although the minimum value is 0 (zero), 25% of the municipality's present values between 0 and 30.97 functional class 3 segments per 1000 meters. The mean and median also present similar values, with the difference of 2 units only, 48.16 and 45.43, respectively.

Furthermore, only 25% of the municipalities have values higher than 61.70 and lower than 278.54 functional class 3 segments per 1000 meters.

From the graph it can be seen that the distribution of functional class 3 per thousand meters is somewhat irregular, although one can notice a large concentration in the north of Portugal, with values between 65.46 and 278.54 segments of functional class 3 per thousand meters.

Statistics		
fun3_p_comp		
N	Valid	278
	Missing	0
Mean		48.1596
Median		45.4326
Std. Deviation		28.86601
Minimum		.00
Maximum		278.54
Percentiles	25	30.9719
	50	45.4326
	75	61.6964

Table 4.11 - Analysis of variable fun3_p_comp

Source: Here

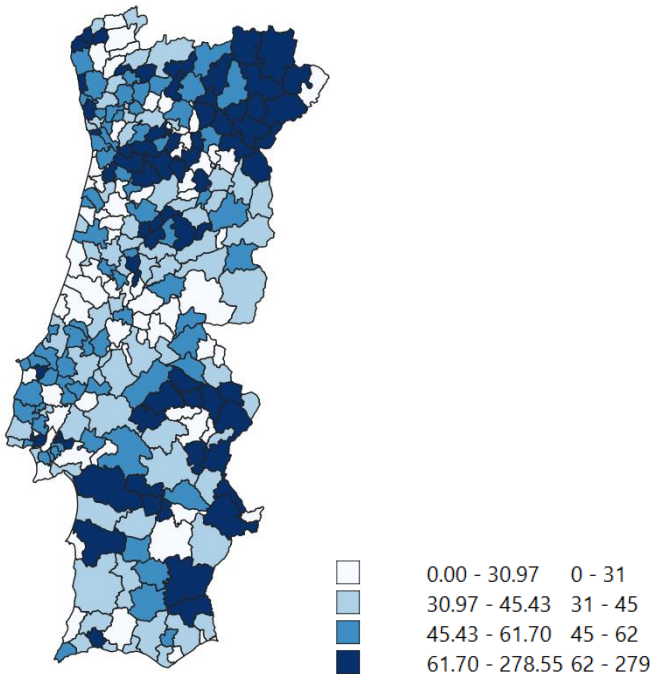


Figure 4.9 - Distribution of variable fun3_p_comp

Source: Here

The functional class 4 represents large traffic volume with moderate speed between neighborhoods. It presents a total of 278 values, which represents the number of municipalities in Continental Portugal. The distribution of this variable, functional class 4 segments per 1000 meters, has a minimum value of 0 (zero) and a maximum value of 295.22, as shown in the table below, Table 4.12.

The mean and median values are close, presenting a difference of only 3 units, as can be seen in the table below, based on the variation of the variable, which is approximately 300 units.

The distribution of the different values by municipalities can be seen in Figure 4.10 showing many municipalities in the center of mainland Portugal, with a low number of segments 4 per 1000 meters, with a range of values from 0 to 71.87 segments of functional class 4 per 1000 meters.

Statistics		
fun4_p_comp		
N	Valid	278
	Missing	0
Mean		104.2378
Median		101.9112
Std. Deviation		40.15510
Minimum		.00
Maximum		295.22
Percentiles	25	78.3744
	50	101.9112
	75	127.0628

Table 4.12 - Analysis of variable fun4_p_comp

Source: Here

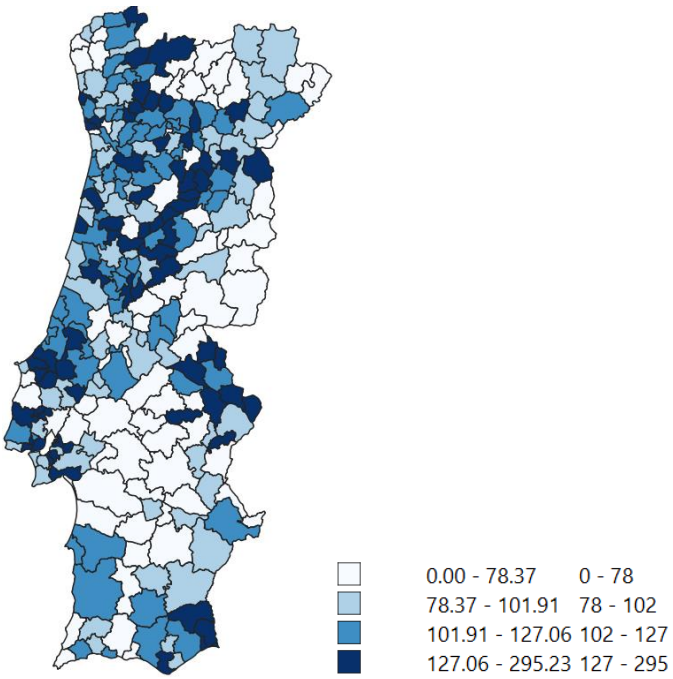


Figure 4.10 - Distribution of variable fun4_p_comp

Source: Here

4.1.5) Velocity

The database contains some variables that represent speed, one of the major causes mentioned in the literature. Thus, it was important to include these variables in the study and associate them with other variables also mentioned in the literature.

The database that contains the variables representing speed, belongs to Here (mentioned above) and contains 8 levels of speed, ranging from 2 to 8 (Here, https://www.here.com/, viewed on 15/08/2022).

It was important to group some speeds in order to help the study and to group them according to Portuguese roads. The speeds were organized in 4 groups (Here, https://www.here.com/, viewed on 15/08/2022):

- Speed 2 and 3: segments have a speed between 91 - 130 km/h;
- Speed 4 and 5: segments with speed between 51 - 90 km/h;
- Speed 6: segments with speed between 31 - 50 km/h;
- Speed 7 and 8: segments with speed between 0 - 30 km/h.

Velocity 6

Speed 6 represents segments with speeds between 31-50km/h. The municipalities with speed 6 segments per thousand meters have a range of 833.01, with 25% of the municipalities having at most 161.22 speed 6 segments per 1000 meters.

Statistics		
vel6_p_comp		
N	Valid	278
	Missing	0
Mean		287.1519
Median		259.7190
Std. Deviation		160.94112
Minimum		.00
Maximum		833.01
Percentiles	25	161.2156
	50	259.7190
	75	392.6016

Table 4.13 - Analysis of Variable vel6_p_comp

Source: Here

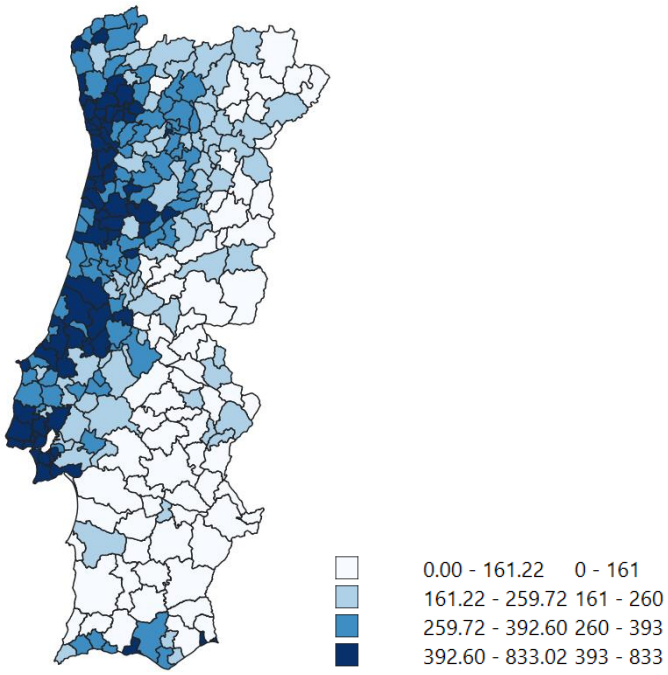


Figure 4.11 - Distribution of the variable vel6_p_comp

Source: Here

Velocity in Urban Areas

Since many accidents are related to urban areas because they are areas where there are many pedestrians (World Health Organization, 2023), it was necessary to create these 3 variables. As expected, zones with a lower speed, namely speed 7 and 8 which comprise speeds between 0 and 30 km/h are the ones with the highest values, 325.05 speed segments 7 and 8 per 1000 meters.

Speeds 2 and 3 encompass speeds between 91 and 130 km/h, accounting for at least 50% of municipalities as 0 (zero) speed segments 2 and 3 in urban areas per 1000 meters, as shown in Table 4.14, showing a maximum of 42.37 in a single municipality, as we can see in Figure 4.12.

The segments with speeds between 51 and 90 km/h are more frequent, 25% of the municipalities have a maximum of 2.11 speed segments 4 and 5 in urban areas per 1000 meters, and 25% of the municipalities also have a minimum value of 7.70.

At lower speeds, there is a significant increase in the number of segments in urban areas per 1000 meters, as seen in speed 7 and 8, Table 4.14, totaling 325.05. As shown in Figure 4.12, the value of the number of segments of the different speeds per 1000 meters increases with the lower speed, there is also a greater number of segments along the coast of mainland Portugal, this is explained by the higher population density along the coast, as you can see in the point 4.1.1) Population, where the largest cities are located, such as Lisbon, Porto, among others.

Statistics				
		vel2e3_urban_p_comp	vel4e5_urban_p_comp	vel7e8_urban_p_comp
N	Valid	278	278	278
	Missing	0	0	0
Mean		1.7082	5.9338	91.6023
Median		.0000	4.1106	78.0694
Std. Deviation		4.89777	6.11842	64.32470
Minimum		.00	.00	.00
Maximum		42.37	43.16	325.05
Percentiles	25	.0000	2.1126	39.8450
	50	.0000	4.1106	78.0694
	75	.8256	7.7041	127.3644

Table 4.14 - Analysis of speed variables in urban areas

Source: Here

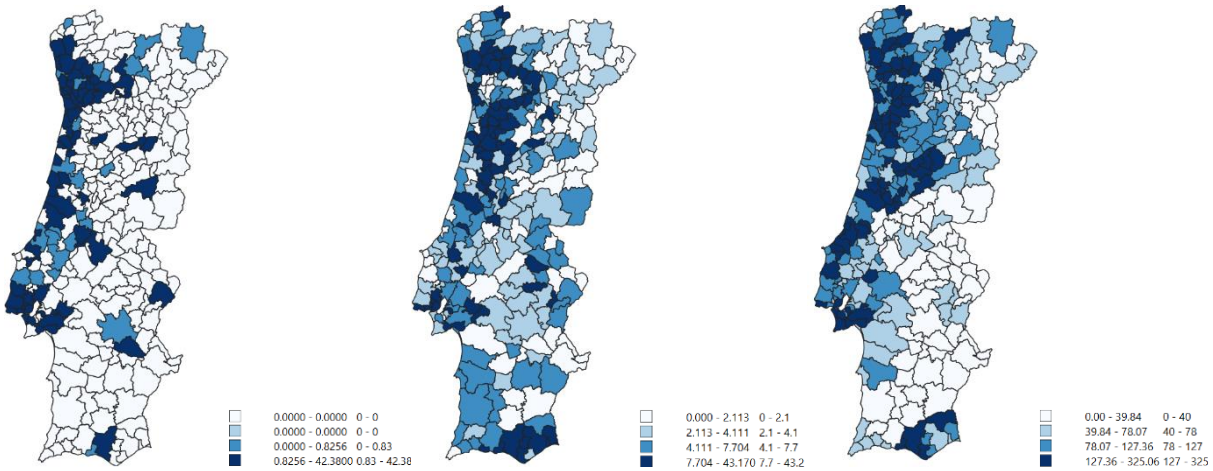


Figure 4.12 - Distribution of variables vel2e3_urban_p_comp, vel4e5_urban_p_comp and vel7e8_urban_p_comp, respectively

Source: Here

Velocity in Tunnels

Luminosity is an external factor with a certain weight in road accidents (Albuquerque et al., 2021), 4 variables were created with tunnels, based on the fact that when entering a tunnel, luminosity varies both for more or less luminosity. Associated with this factor, the speed was added being another determining factor in road accidents.

Analyzing Table 4.15, it can be seen that there are few tunnels along the Portuguese territory since in all variables with different speeds, at least 75% of the municipalities show a value equal to 0 tunnels with different speeds per 1000 meters.

However, the higher the speed, the higher the value of tunnels at a certain speed per 1000 meters, with a maximum of 7.29 at speed 2 and 3 and 0.42 at speed 7 and 8.

Statistics					
		vel2e3_tunel_p_comp	vel4e5_tunel_p_comp	vel6_tunel_p_comp	vel7e8_tunel_p_comp
N	Valid	278	278	278	278
	Missing	0	0	0	0
Mean		.0887	.0327	.0318	.0029
Median		.0000	.0000	.0000	.0000
Std. Deviation		.52947	.37481	.25828	.02819
Minimum		.00	.00	.00	.00
Maximum		7.29	6.13	3.47	.42
Percentiles	25	.0000	.0000	.0000	.0000
	50	.0000	.0000	.0000	.0000
	75	.0000	.0000	.0000	.0000

Table 4.15 - Analysis of speed variables in tunnels

Source: Here

4.1.6) Intersections

The road conditions, like intersections are also a relevant factor in road accidents (Alkan et al., 2021)

The number of intersections per 1000 meters presents a range of 10.55, with only 25% presenting values higher than 1.38 intersections per 1000 meters. The intersections variable presents a mean and a median of 1.15 and 0.91 intersections per 1000 meters, respectively.

With the analysis of the Figure 4.13 it appears that there are different values of intersections per 1000 meters throughout Continental Portugal.

Statistics		
intersecoes_p_comp		
N	Valid	278
	Missing	0
Mean		1.1528
Median		.9069
Std. Deviation		1.16323
Minimum		.00
Maximum		10.55
Percentiles	25	.5204
	50	.9069
	75	1.3847

Table 4.16 - Analysis of the variable intersecoes_p_comp

Source: Here

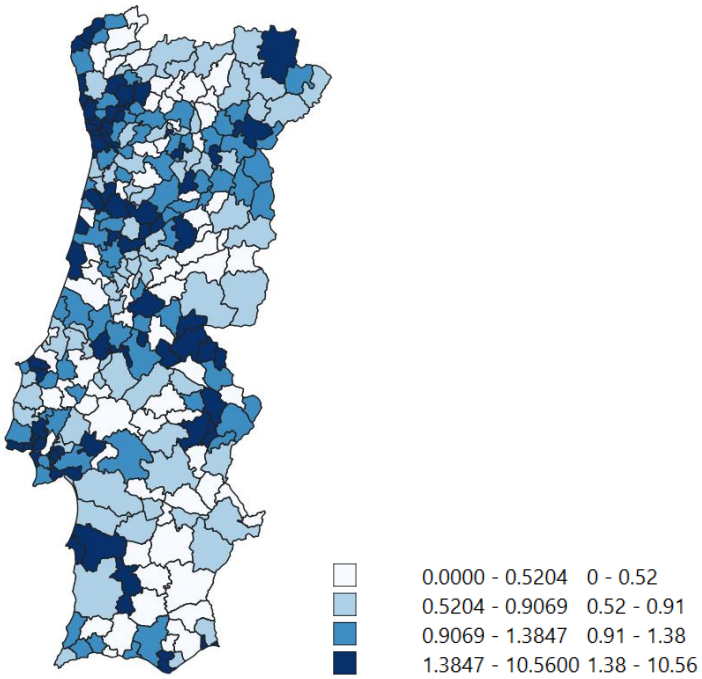


Figure 4.13 - Distribution of the variable intersecoes_p_comp

Source: Here

4.1.7) Motorcyclist and Pedestrians

According to World Health Organization more than 50% of all road traffic death are Motorcyclists and pedestrians (World Health Organization, 2022) are two big weights in accidents and accident severity, so two variables were created that combine motorists, all four-wheeled vehicles along with motorcyclists, and pedestrians and motorists.

As most of the roads where four-wheelers pass motorcycles, there is a minimum of 1438.65 car and motorcycle segments per 1000 meters in a municipality and a maximum of 2000 car and motorcycle segments per 1000 meters in a municipality. On the contrary, the roads where pedestrians and road users can pass are lower, with a maximum of 1998.49 pedestrian and road user segments per 1000 meters in a municipality and a minimum of 687.18.

Statistics			
		auto_moto_p_comp	ped_rodo_p_comp
N	Valid	278	278
	Missing	0	0
Mean		1966.7613	1016.8972
Median		1980.0790	952.5250
Std. Deviation		49.87992	267.70341
Minimum		1438.65	687.18
Maximum		2000.00	1998.49
Percentiles	25	1957.3242	920.6675
	50	1980.0790	952.5250
	75	1991.4813	980.7223

Table 4.17 - Analysis of pedestrian and motorcycle variables

Source: Here

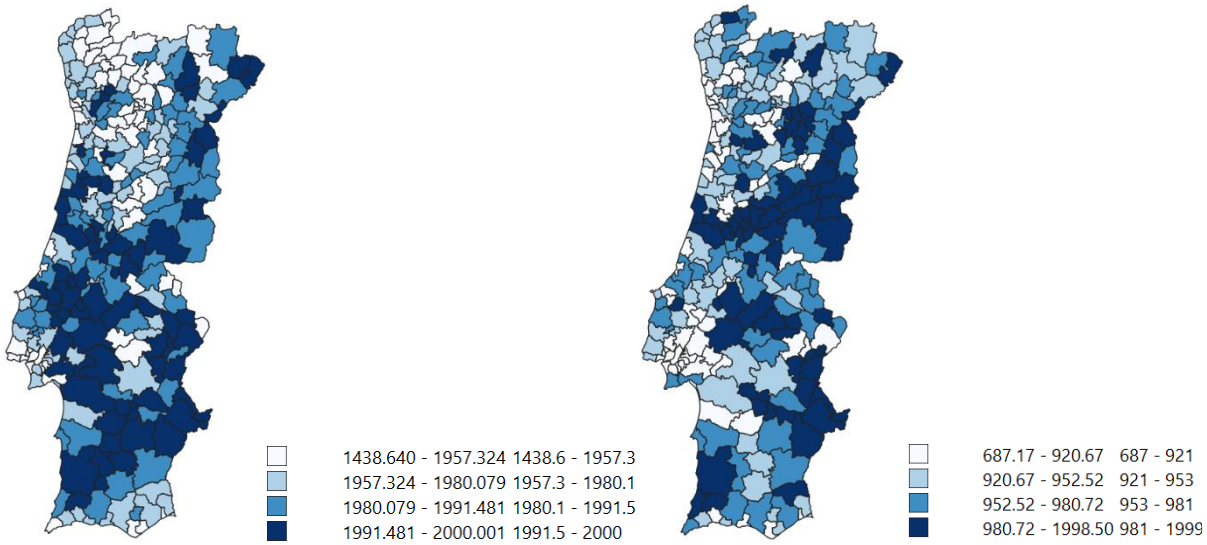


Figure 4.14 - Distribution of motorcycle and pedestrian variables, respectively

Source: Here

4.1.8) Accidents

As the objective of this work is to find out the weight that external factors have on accidents, the total number of accidents per length (meters) was considered as the dependent variable. From Figure 4.15 it is clearly noticeable that accidents occur mainly in the large arable areas of Lisbon

and Porto. Accidents have a more accentuated value along the coastline and decrease when moving inland.

The maximum value of accidents per 1000 meters is only 1.76, and it is important to note that there are municipalities with 0 (zero) accidents per 1000 meters and, at least, 75% of the municipalities have values lower than 1 accident per 1000 meters.

Statistics		
tot_ac_p_comp		
N	Valid	278
	Missing	0
Mean		.1217
Median		.0591
Std. Deviation		.19573
Minimum		.00
Maximum		1.76
Percentiles	25	.0292
	50	.0591
	75	.1274

Table 4.18 - Analysis of tot_ac_p_comp variable

Source: INE

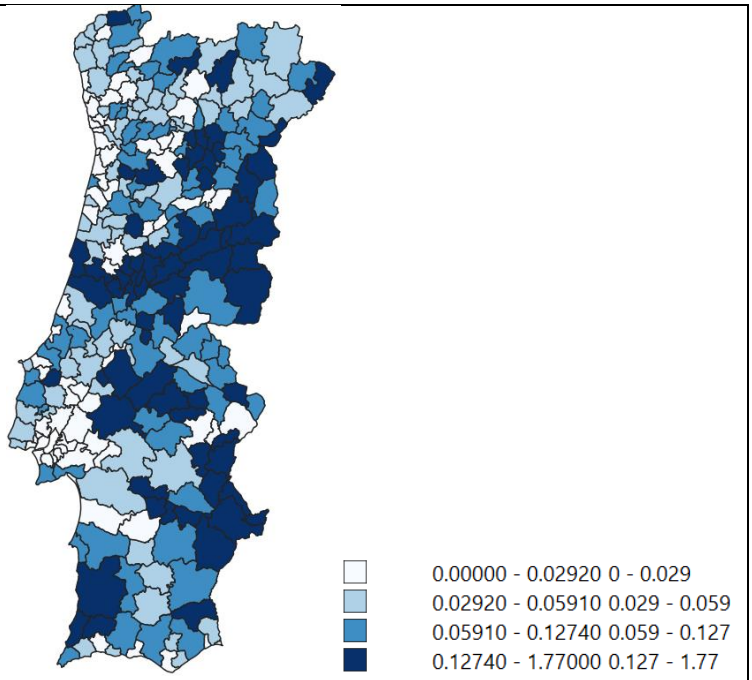


Figure 4.15 - Distribution of the variable tot_ac_p_comp

Source: INE

4.1.9) Victims

The dependent variable associated with victims was created from the variable corresponding to total victims (light, serious and fatal victims) per 1000 inhabitants. Furthermore, it will be important to study that, although accidents and victims are related in some way, whether on a statistical level they are related.

Since we will be conducting the study of which and with what weight external factors have on victims, that's how much external factors influence the likelihood of being a victim, so this variable will also be related against external factors, as a dependent variable.

Contrary to accidents, the likelihood to be a victim per 1000 citizens is not higher in the large areas of Lisbon and Porto, there are more victims in the central and southern areas, but analyzing the Table 4.19 and the Figure 4.16 referring to the total number of victims, there is a greater number

of victims along the coast, mainly in big cities. The maximum value is 12 victims per 1000 citizens, but in the real number is 3418 victims in just one municipality (Lisbon).

Statistics		
tot_vit_p_pop		
N	Valid	278
	Missing	0
Mean		4.5586
Median		4.3906
Std. Deviation		1.49434
Minimum		.00
Maximum		12.00
Percentiles	25	3.5973
	50	4.3906
	75	5.2838

Statistics		
Totaldevitimas		
N	Valid	278
	Missing	0
Mean		158.2914
Median		64.0000
Std. Deviation		280.23676
Minimum		.00
Maximum		3418.00
Percentiles	25	30.0000
	50	64.0000
	75	182.2500

Table 4.19 - Analysis of the variable tot_vit_p_pop and Totaldevitimas

Source: INE

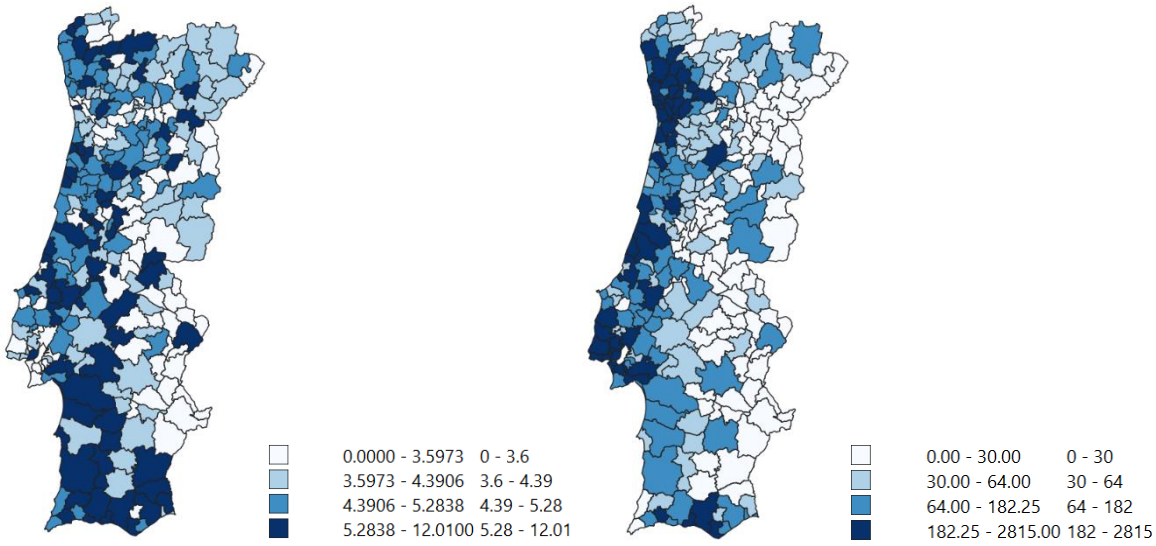


Figure 4.16 - Distribution of the variable tot_vit_p_pop and total of victims in mainland Portugal

Source: INE

4.2) Relation between dependent variables

After the analysis of all variables, it is important to prepare and understand which variables can really be significant for the models.

Since accidents cause victims, we felt the need to verify if the external factors that influence accidents would be the same as those that influence victims, and what is the weight of these, since the purpose is to verify if and what influence external factors have on accidents and victims, these have to be dependent on external factors. So, the variables relative to accidents per 1000 meters

and victims per 1000 inhabitants are the dependent variables and all of variables relative a external factors are an independent variables.

As mentioned above, accidents cause victims and to do correctly specify the model structure and analytical approach to follow, the first step has to do with the casual link between accidents and victims.

Firstly, it is important to analyze the variables accidents and victims with their absolute values. For this analysis we did a correlation analysis (annex C), a scatter plot analysis (annex D) and a linear regression (equation 2) as a vehicle to explore causality between accidents and victims. With this analysis, it was verified that accidents have a strong correlation with victims, presenting a Pearson Correlation value of 0.99, practically 1 (meaning that these variables levels vary almost in the same proportions). Given the previous result, the scatter plot shows naturally almost perfect line.

Additionally, as a vehicle to test the relation between the variables, equation 2 was estimated.

$$victims_i = \beta_0 + \beta_1\, accidents_i + u_i \quad , \qquad (2)$$

Where

$victims_i$= total number of victims in municipality I,

$accidents_i$ = total number of accidents in municipality I,

u_i = the error term.

The estimate of equation 2 confirmed the literature and the common sound, that is, the more accidents, the more victims. With the analysis of the equation, we can say, according to the absolute data taken from the INE, for each accident there is an increase in victims by 1.229 (β_1). Furthermore, equation 2 has a constant value of 6.978 (β_0) and t a value of 8.5 and 387.5 for β_0 and β_1, respectively.

However, as stated previously, once we are dealing with spatial data, beside having to compare relative measures, our variable of interest is more the likelihood to have an accident per spatial unit, rather than the total number of accidents. So, the accidents per 1000 meters is a proxy for the likelihood of having an accident in the geographic space and the victims per 1000 inhabitants is a proxy for the likelihood of being a victim in a determinate location.

In order to analyze the relationship between these two proxy variables (and candidate dependent variables) for the likelihood for accidents and victims, it was necessary to analyze the

linear correlation between them (annex F), analysis of scatter plot (annex G) and the estimate of an equation 3.

Accidents per 1000 meters and victims per 1000 inhabitants don't show a strong linear correlation, as they have a Sig of only 0.554 (annex E) and a R square equal to 0.001. In addition, it is possible to see in annex E (scatter plot) the values are all dispersed.

Additionally, to test the existence of any causal link between the variables, equation 3 was performed.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 {x_i}^2 + \beta_3 {x_i}^3 + \beta_4 W x_i + u_i \quad (3)$$

Where

y_i= victims per 1000 inhabitants,

x_i = accidents per 1000 meters,

W = queen contiguity matrix,

u_i = the error term.

Equation 3 removed any probability of the existence of any causality between the variable representing the proxy for the likelihood of accidents and victims. For this conclusion, a hypothesis test was created:

$$H0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

$$H1: there\ is\ \ at\ least\ one\ \beta_i \neq 0$$

To teste the previous hypothesis, analyzing the F test (F=1.538, sig=0.191), we don't have statistical evidence in order to reject $H0$ (even at the 0.15 significance level), meaning that (at least for the tested structure), we have no sign that there might exist a causal link between the variables. This is an important conclusion once it conditions the modelling strategy followed. Once both variables are independent, it was decided to treat them independently, that is, to do a model to accidents and other model to victims against external factors and to understand which factors and what is the weight of each external factor in both victims and accidents.

4.3) Spatial Dependence

Since we are dealing with spatial information, it is important to understand if there is spatial dependence or not. To analyze spatial dependence, we used the Moran Index.

The Moran's index is a measure of spatial correlation (Chen, 2021) to identify a positive spatial relationship between accidents and the surrounding factors, measure (Oetomo et al., 2017), in this case the external factors.

According to García (2020) the Moran index was created by Patrick Alfred Pierce Moran in the early twentieth century in order to calculate spatial autocorrelation, in this case for the dependent variables. The Moran index is calculated as follows:

Source: García (2020)

$$I = \frac{N}{\sum_i \sum_i w_{ij}} \frac{\sum_i \sum_i w_{ij} (X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2} \quad (4)$$

Where:

N – Number of spatial units indexed by i and j;

X – independent variable;

$\bar{X}$ – Arithmetic mean of variable X;

w_{ij} – Spatial weight matrix element.

Analyzing Moran's Index of the two dependent variables, it was found a higher spatial correlation in the variable tot_ac_p_comp than in the variable tot_vit_p_pop, corresponding to a value of 0.482 and 0.147, respectively, meaning that accidents have higher spatial dependence than victims (annex H).

In order to analyze spatial dependence, the contiguity matrix was created using the Queen criteria. The Queen criteria means that everything around the observation will be influenced by these observations, in this case the accidents in a municipality influenced other accidents in others municipalities thar are beside, as shown in Figure 4.17.

	UnitC1	UnitB2	UnitC2	
	UnitB1	UnitA	UnitB3	
	UnitC4	UnitB4	UnitC3	

Figure 4.17 – Queen Contiguity

Source: Oetomo, H. W, et. al. (2017)

According to Chen (2021) The weights express the neighbor structure between the observations as a $n \times n$ matrix **W** in which the elements $W_{ij}$of the matrix are the spatial weights.

Source: Chen (2021)

$$W = \begin{bmatrix} W_{11} & W_{12} & \dots & W_{1n} \\ W_{21} & W_{22} & \dots & W_{2n} \\ W_{n1} & W_{n2} & \dots & W_{nn} \end{bmatrix}$$

The value of the weights is never zero when i and j are neighbors, and when i and j are equal, W_{ij} will have a value of 0 (zero) (Geoda, 2022).

After calculating the contiguity matrix (W), the spatial lags of the dependent variables, tot_ac_p_comp and tot_vit_p_pop, were calculated. According to Oetomo, et al. (2017), the most important in spatial analysis is the contiguity matrix with the values referent to spatial weights, that determine the weights between locations.

Additionally, although the number of variables had already been reduced, it would be necessary to understand which variables could really bring some knowledge to the final model, thus, several models were created and analyzed with de linear regression like de equation 5.

$$y_i = ax_{1i} + bx_{2i} + cx_{3i} + \cdots + zx_{ni} + \beta W x_{ki} + u_i \qquad (5)$$

Being,

y_i – dependent variable (accidents and victims),

x_k – independent variables, with k=1, 2,….n,

W – the queen contiguity matrix,

u_i- the error term.

“In a regression context, spatial effects pertain to two categories of specifications. One deals with spatial dependence, or its weaker expression, spatial autocorrelation, and the other with spatial heterogeneity.” (Anselin, 2003)

The models created in order to verify which variables could bring a certain knowledge and results to the final models, are linear regressions with spatial lag component

In addition, it is necessary to pay attention to the factors mentioned in a paragraph above: multicollinearity, spatial heterogeneity e R^2.

Following a general-to-specific (gts) significance approach, the variables that did not present significant coefficients at the 0.05 level were removed, obtaining a model with 6 variables for the accidents as shown in annex I and the equation 6.

$$\boldsymbol{Tot\_a\_comp_i = \beta_0 + \beta_1 vel6_{comp_i} + \beta_2 vel6_{tu_i} + \beta_3 inter_{comp_i} + \beta_4 fun1_{comp_i} + \beta_5 vel4e5_{tu_i} + \beta_6 Popul_{2018_i} + \beta_7 W_{tot_{a_{comp_i}}} + u_i} \qquad (6)$$

With the variables defined and prepared for modeling, a linear regression model was obtained with Spatial Lag with an explanatory capacity of 70.27%. All variables have a positive $\boldsymbol{\beta}$, that is, they all positively influence accidents, contribute to more accidents. Among all variables, speeds between 31-90 km/h are the variables that most influence accidents. In the other hand, the 2018 population and speed segments between 31-50 km/h are the variables that least positively influence accidents (annex I).

In relation to the victims variable, the same procedures were also performed as in the accident models, and thus a model with 4 variables was obtained, as can be seen in the annex J and equation 7.

$$\boldsymbol{Tot\_v\_pop_i = \beta_0 + \beta_1 Popul_{2018_i} + \beta_2 fun1_{comp_i} + \beta_3 vel6_{tu_i} + \beta_4 ampli\_tura_i + \beta_6 Popul_{2018_i} + \beta_5 W_{tot_{v_{pop_i}}} + + u_i} \quad (7)$$

From the equation an explanatory capacity of 12.45% was obtained, where the population of 2018 and the temperature amplitude negatively influence the victims, that is, when the variables increase by 1 unit, the victims decrease. On the contrary, the variable that most positively influences (1.17) for the increase in victims are speeds 31-50 km/h in tunnels.

Chapter 5 – Results

After the analysis of all variables, the necessary modifications and their selection, the modeling creation phase began, from which the final results and conclusions will emerge.

Like we saw in the chapter 4, the dependents variables do not saw a large linear correlation, the variable corresponding to accidents are a higher value to spatial correlation, furthermore with the linear regressions addressed in the chapter 4, we saw that variables on model accidents are different that variables on model severity, so the modeling part was divided into 2 important sections since it will be state 2 dependent variables and that these have only a moderate correlation between them at the statistical level, as can be seen in the data preparation chapter.

Thus, the modeling chapter will also be divided in two sections: accidents and victims. The weight, w, discussed in the previous section, data preparation section is an important variable and included in all models, this variable (spatial lag) was created by the value of the weights for each municipality multiplied by the value of the dependent variable, in this case, tot_ac_p_comp or tot_vit_p_pop.

Modeling was carried out in two steps. First, we defined the structure of the models, and then we applied a simulation (bootstrapping) study for these models using the same structure, that is, generating new data from the victims and accidents databases.

According to Efron and Tibshirani (1986) the bootstrap is a methodology based in a computer method which substitutes a large amount of computation for simplify the study and increasingly a good data analytic.

The bootstrap method is a method that can be applied to databases with a large number of data, which consists in generating new data based on two principles, the non-parametrized method that generates data randomly and the parameterized method that generates new data based on the F distribution in order to normalize the distribution of variables (Wehrens et al., 2000), like we used. So in this case we use the bootstrap method to create a simulation study.

The bootstrap method and simulation study was used with 4 objectives: eliminate potential extreme effects and isolated cases (outliers) that are specific to Portugal; know the effects and weights asymptotically; potentially generalize to other countries with a similar road structures; potentially generalize to Portugal for new future road structures.

Thus, with the sample data, we identified the structure of the models. Then we went to do a simulation study, where we applied the structure model used in sample data, to know the weights

and explanatory capacity asymptotically, that is we randomly generated new 300,000 municipalities, to get closer to reality. So, the evaluation refers to 300,000 data.

5.1) Evaluation metrics used

To draw conclusions from the models, it is essential to evaluate them, so the evaluation metrics are the predictors of importance and the accuracy.

Predictor's importance

Predictors' importance is the center for analyzing the weight that external factors have on accidents and victims.

"Predictor importance can be determined by computing the reduction in variance of the target attributable to each predictor, via a sensitivity analysis." (IBM, 2021)

The predictors are ranked according to a sensitivity measure that is calculated as shown in equation 8.

Source: IBM, 2021

$$S_i = \frac{V_i}{V(Y)} = \frac{V(E(Y|X_i)}{V(Y)} \quad (8)$$

$V(Y) -$ unconditional variance of the output (independent variable)

$V\big(E(Y|X_i)\big) -$ variance of the integral of variable y over x

Subsequently, it is necessary to normalize the sensitivity measure, which is performed as follows, as presented in equation 9.

Source: IBM, 2021

$$V\ I_i\ = \frac{S_i}{\sum_{j=1}^{k} S_j} \quad (9)$$

Accuracy

There are several metrics to evaluate the effectiveness of different models, accuracy being the most widely used, in classification models, representing the percentage of correctness in the test predictions (Hasheminejad et al., 2017).

$$Accuracy = ESS/TSS, \quad (10)$$

$$ESS = \sum(\hat{y} - \bar{y})^2, \quad (11)$$

$$TSS = \sum(y_i - \bar{y})^2 , \quad (12)$$

Where:

ESS = Explain sum of squares,

$TSS =$ Total of squares.

y_i = dependent variables,

$\bar{y}$ = average of dependent variables

$\hat{y}$ = predicted value of y given x

As explained above (chapter 4) the relative dependent variables were treated independently. Thus, results were presented for the models of accidents per 1000 meters and results for the models of victims per 1000 inhabitants.

5.2) Accidents

For the accident variable a number of different modeling forms were created, ranging from neural networks, decision trees, and regressions, approaches talked about in the literature review chapter 1.

The models generated were all taken into account and some options were tweaked in order to avoid overfitting, in this case the neural network has only 1 hidden layer, and in the decision trees it was restricted to, at most, the tree could grow 7 layers and with a maximum of 5 child nodes or 5% of the data in the child node and 10% in the parent nodes. Additionally, exist models with components bagging and boosting, to improve de measure accuracy and avoid overfitting. In the regressions the models used a linear regression strategy.

All models were run with the variables selected, and so the input variables are: Popul_2018, vel6_comp, vell6_tu, inter_comp, fun1_comp, vel4e5_t e w_acid (weight), as mentioned in the data preparation chapter.

The models created have an accuracy range between 55% and 79%, and the model with the lowest accuracy is represented by the C&RT decision tree with the Bagging particularity, that is, it performs a bootstrapping and performing several models in order to avoid overfitting, and the model with the best accuracy rate is the Neural Network with 79.9%, as can be seen in Table 5.1.

Type of Model	Accuracy
Neural Network	0.799
Regression	0.705
Random Tree	0.714
Cart Boosting	0.64
Cart Bagging	0.556
LSVM	0.644
SVM	0.794
Average	0.6931

Table 5.1 - Accidents Models: Accuracy

Regarding the variables included in the models, each one has its importance, on average the variable with the highest value in the predictors is the w_acid, that is the spatial lag , since there is a spatial dependence, it was already expected that the weight is one of the variables with the highest weight. On the contrary the variable with less weight is vel_6, stretches with speeds between 31-50 km/h, with values of 28% and 7%, respectively (Table 5.2).

Variables	Average Importance
w_acid	0.2814
vel6_tu	0.1429
vel4e5_tu	0.1629
inter_comp	0.1129
popu_2018	0.1214
fun1_comp	0.0957
vel6_comp	0.0714

Table 5.2 - Average Importance: Variables

The importance of external factors varies according to the model under study, in the case of sections with functional class 1 (allows a large volume of traffic movement at maximum speed), it has a greater importance in the Random Tree model, representing 14.28% (0.714*0.2) of road accidents. The intersections variable has no weight in the SVM model, and its greatest representation is in the LSVM model, explaining 12.88% of accidents, according to this model.

As the variable representing the intersections the population also shows no explanatory value in the SVM model, having a higher explanatory value in the LSVM model.

In the C&RT Boosting and Bagging models all variables have an explanatory value of 0.15, representing 9.6% and 8.34%, respectively, of the road accidents. On the contrary, the variable

represented by functional class 1 presents an importance of only 0.07 in the C&RT Boosting and Bagging models.

Additionally, the speed 6 in tunnels has a higher explanatory value in the SVM model having a weight of 0.22 in the model. The neural network has the highest value of accuracy and for this model the most explanatory variable is the 2018 population of mainland Portugal representing 0.2, i.e., the 2018 population is able to represent 15.58% of road accidents, when using the neural network model (annex K).

5.3) Victims

The creation of the models for the victims were all based on the variables chosen with the help of the literature review and the regressions that were analyzed in Geoda, so the variables that served as input are: popul_2018, vel6_tu, fun1_comp, ampli_tura, w_vit and as target tot_v_pop. The models used were also created based on the literature and in this way the different algorithms were selected: neural network, decision trees, LSVM and regression with rule to avoid overfitting and improve the accuracy, with components boosting and bagging.

After creating all models with the input variables popul_2018, vel6_tu, fun1_comp, ampli_tura, w_vit and target, tot_v_pop and a bootstrapping of 300,000 observations, accuracies range between 11.2% and 15.5%, representing the CHAID Boosting and Random Tree algorithm, respectively, were obtained, as shown in annex K.

The neural network and regression algorithms show above average (12.48%) accuracy values of 13.7% and 12.9%, respectively. All other models present values below 12.48%.

Type of Model	**Accuracy**
Neural Network	0.137
Regression	0.129
Random Tree	0.155
Cart	0.105
Cart Boosting	0.122
Cart Bagging	0.111
Chaid Boosting	0.112
Chaid Bagging	0.102
LSVM	0.127
Average	0.1222

Table 5.3 - Victims Models: Accuracy

The factor with the highest explanatory value, on average, is the variable representing the weight calculated by the contiguity matrix with the target variable, presenting a value of 28.9% and, on its opposite, the factor that has the least explanatory value is tunnels with speed 6 (31-50 km/h) corresponding to only 9.4%.

Within the external factors, the factors with the highest explanatory capacity are the temperature range and the sections represented by functional class 1, having an explanatory capacity of 24.7% and 21.4%, respectively, as can be seen in Table 5.4.

Variables	**Average Importance**
w_vit	0.298
amplit_termica	0.247
fun1_comp	0.214
popu_2018	0.158
vel6_tu	0.094

Table 5.4 - Victims Models: Predictors of Importance

The value of the explanatory values can vary between 0 and 1. Moreover they explain only a part, that which belongs to the accuracy of the model we will address. In this way the temperature range is an explanatory variable, capable of explaining 3.3% of the victims in the C&RT model and 2.3% in the neural network created.

Regarding the sections with functional class 1 it is an explanatory variable with an importance of 29% representing an explanatory capacity of 3.2% of the victims resulting from accidents. Regarding the variable referring to the 2018 Population, three of all the models created this variable presents an importance lower than or equal to 5%, more specifically in Regression, LSVM, CHAID Bagging and C&RT, on the contrary, in the remaining models, Neural Network, Random Tree, CHAID Boosting and C&RT Boosting and Bagging, the variable presents an importance of at least 23%.

Finally, when analyzing the variable representing tunnels with speeds between 31-50km/h it was found that this variable has no weight in the CHAID and C&RT models with the Boosting specificity (annex L).

5.4) External factors explanatory capacity

Based on accidents, the models analyzed have an explanatory capacity of 69.3%, on average, this mean that external factors can explain a part of accidents. With all the analysis of the literature review, the variables and their selection, we arrived with a total of 6 variables capable to explain part of the accidents. These variables are: velocity between 31-50 km/h in tunnels, velocity between 51-90km/h in tunnels, velocity between 31-50 km/h on the road, intersections, population in 2018 and roads allows a large volume of traffic movement at maximum speed.

Within accidents models, the variable with the highest importance value is velocity between 51-90km/h in tunnels, and the variable with the lowest importance value is velocity between 31-50 km/h on the road, with 16.29% and 7.14%, on average, respectively.

As already mentioned above, in chapter 2, there are many other factors, namely intrinsic factors, with a greater weight, so it is acceptable and normally for the variable w_acid to be the variable that presents a greater importance, on average, both in accidents and in victims, presenting values of 28.14% and 29.14% respectively.

Based on victims, the models have an explanatory capacity of 12.22%, on average, that is a lowest value that accidents models. After all analysis described throughout this study, the victims models have 4 variables capable to explain part of the victims. The variables are: population in 2018, average of temperature, velocity between 31-50 km/h in tunnels and roads allows a large volume of traffic movement at maximum speed.

In these models we can see that the variable with a greater importance value is the average of temperature with, on average, 24.7% and the variable with a lowest importance value is velocity between 31-50 km/h in tunnels with, on average, 9.4%.

So based on these models we can suggest that external factors can influence the accidents and the victims. On average, the explanatory capacity of each variable, in accidents models are between 11.3% and 4.9%, as it is represented in the Table 5.5, with the variable velocity between 51-90 km/h in tunnels and velocity between 31-50 km/h, respectively.

The explanatory variable with the highest explanatory power on average (vel4e5_tu) has a higher explanatory power of 15.7% with the Random Tree model and the lower explanatory power of 7.7% in the LSVM model. The explanatory variable that have a lowest explanatory capacity is vel6_comp with an explanatory capacity of 0.8% with de SVM model.

Variables	Explanatory Capacity
vel6_tu	9.9%
vel4e5_tu	11.3%
inter_comp	7.8%
popu_2018	8.4%
fun1_comp	6.6%
vel6_comp	4.9%

Table 5.5 – Explanatory Capacity: Accidents Models

Additionally, based on de victims models the variable with higher explanatory capacity average temperature with 3% and lowest explanatory capacity is velocity between 31-50 km/h in tunnels with 1.1% (Table 5.6).

The amplit_termica is the variable with the higher explanatory capacity on victims, with 3%, on average. This variable has an explanatory capacity of 2.3% in Neural Network (the lowest value) and the 3.5% explanatory capacity in Regression model (the highest value).

Variables	Explanatory Capacity
amplit_termica	3.0%
fun1_comp	2.6%
popu_2018	1.9%
vel6_tu	1.1%

Table 5.6 – Explanatory Capacity: Victims Models

Regarding the explanatory variable of spatial dependence, this is represented by external and internal factors not accounted for in these models. Thus accidents, the external factors have an explanatory capacity that varies between 38.9% (SVM) and 55.7% (Random Tree).

Type of Model	Explained Total	Explained w_accidents	Explained external factors
Neural Network	79.9%	24.8%	55.1%
Regression	70.5%	27.5%	43.0%
Random Tree	71.4%	15.7%	55.7%
Cart Boosting	64.0%	9.6%	54.4%
Cart Bagging	55.6%	8.3%	47.3%
LSVM	64.4%	15.5%	48.9%
SVM	79.4%	40.5%	38.9%
Average	69.31%	20.27%	49.04%

Table 5.7 – Explanatory Capacity: External factors (accidents)

Regarding the models referring to victims, it appears that the explanatory capacity of the models is much lower, when compared to accidents, varying between 10% (Chaid Bagging and Cart) and 15% (Random Tree), taking into account the variables responsible for dependence spatial is included. This low percentage can be explained by the fact that this proxy likelihood of victims is not explained by the factors inherent to these models. In this way, internal factors are factor with much weight when it comes to victims, compared with accidents. The factor not accounted for in these models have an explanatory power of 90%, on average.

Type of Model	**Explained Total**	**Explained w_victims**	**Explained external factors**
Neural Network	13.7%	2.9%	10.8%
Regression	12.9%	4.5%	8.4%
Random Tree	15.5%	2.2%	13.3%
Cart	10.5%	5.1%	5.4%
Cart Boosting	12.2%	3.1%	9.1%
Cart Bagging	11.1%	2.6%	8.5%
Chaid Boosting	11.2%	2.9%	8.3%
Chaid Bagging	10.2%	3.0%	7.2%
LSVM	12.7%	4.8%	7.9%
Average	12.2%	3.46%	8.7%

Table 5.8 – Explanatory Capacity: Other factors (victims)

Chapter 6 - Conclusion

This research was positioned in a gap that existed in the literature and in studies of road accidents, i.e., existing studies and research focus on specific road sections with specific characteristics and this is replicated in several studies for different countries as we can see in the literature but these studies have a micro approach So, we felt the need to create a macro approach to totality, a profile of the road structure in the Portuguese space and mobility database, that's it how much people move around the municipality.

This research encountered several limitations, mainly finding data that was felt to be needed and used in other studies addressed in the literature review namely accident times, road conditions, and weather accuracy are some examples of data addressed in other studies that were difficult to find or even impossible As the purpose was a macro approach, this prevented the use of intrinsic factors to drivers, as this would require an exhaustive survey of all existing accidents. This non-inclusion of these and other factors may represent a omitted variables problem.

In addition, the fact that it was a study that was little addressed, the literature review was relatively scarce and difficult to find, with everything was a study where it was necessary to learn new concepts, namely being a geospatial problem. Additionally, conducting a study with the purpose of obtaining one or several models capable of responding to a general level of Continental Portugal is a great challenge, because each municipality has its own characteristics and this must be taken into account.

With the four objectives presenting in chapter 5: eliminate potential extreme effects and isolated cases (outliers) that are specific to Portugal; know the effects and weights asymptotically; potentially generalize to other countries with a similar road structures; potentially generalize to Portugal for new future road structures, the simulation study applied to all model structures was found that external factors have some weight in accidents, but are not their main reason, since all models, both in accidents and in victims present an explanatory capacity of 49.04%, on average, by external factors, and the 50%, proximally, are explained by other external factors not addressed or not exemplified in the collected data, or internal, mainly the driver, which, as we have seen, in several studies is pointed out as the main reason.

With all these analyzes and models, it can be said that the external factors, addressed and included in models discussed in chapter 5, can explain part of the accidents in 48%, approximately, and in victims 8.6%, on average.

In accidents the variable velocity between 31-50 km/h in tunnels can explain 9.9%, the velocity between 51-90km/h in tunnels explain 11.3%, velocity between 31-50 km/h on the road explain 4.9%, intersections explain 7.8%, population in 2018 can explain 8.4% and roads allows a large volume of traffic movement at maximum speed explain 6.6%.

The victims does not a explanatory capacity by external factors, the only 8.7% is explained by external factors include in models. The model with de highest explanatory capacity is the Rando Tree with 13.3%, and the lowest explanatory capacity is Cart with 5.4%

In victims the variable average temperature explains 3%, the variable roads allows a large volume of traffic movement at maximum speed explain 2.6%, the population in 2018 can explain 1.9% and velocity between 31-50 km/h explain 1.1%.

The results presented above are in agreement with the literature presented in chapter 2, since most of the causes are represented by other factors than external, such as, for example, factors related to the driver or even the vehicle, or other external factors not accounted in this study.

The major conclusion of this study is that external factors seem to explain a significant part of the likelihood of accidents in space, and not the likelihood of people to be accident victims. Additionally, spatial dependence is more evident in accidents per 1000 meters, and thus, should be taken into account.

For future investigations the models can be complemented with micro data on accidents (more specific) combined with external and internal factors.

References

Albuquerque, V., Oliveira, A., Barbosa, J. L., Rodrigues, R. S., Andrade, F., Dias, M. S. & Ferreira, J.C. (2021). Smart Cities: Data-Driven Solutions to Understand Disruptive Problems in Transportation. *Energies 2021, 14*, 1-25. https://doi.org/10.3390/en14113044

Alkan, G., Farrow, R., Liu, H., Moore, C., Keung, H., Ng, T., Stokes, L. Xu, Y., Xu, Z. Yan, Y & Zhong, Y. (2021) Predictive Modeling of Maximum Injury and Potencial Economic Cost in a Car Accident Based on the General Estimates System Data. *Computational Statistics. 36*, 1561-2575. https://doi.org/10.1007/s00180-021-01074-7

Ameen, J. R. M. & Naji, J. A. (2020). Causal Models for road accident fatalities in Yemen. *Accident Analysis and Prevention, 33*, 547-561.https://doi.org/10.1016/S0001-4575(00)00069-5

Anselin L. (2005). *Exploring Spatial Data with GeoDa™: A Workbook*. Center for Spatially Integrated Social Science (version 2005)

Anselin, L. (2003). Chapter Fourteen Spatial Econometrics. Em B. H. Baltagi (Ed.), *A Companion to Theoretical Econometrics* (pp. 310 – 330). Blackwell Publishing Ltd.

ASF – Autoridade de Supervisão de Seguros e Fundos de Pensões. (2022, June 30). *Estatísticas de Seguros*. https://www.asf.com.pt/NR/exeres/34CBFBFE-40B5-4ECF-AA75-5934E13A57E4.htm

Casado-Sanz, N. Guirao, B. & Attard (2020). Analysis of the Risk Factors Affecting the Severity of Traffic Accidents on Spanish Crosstown Road: The Driver's Perspective. *Sustainability, 12*, 1-26. https://doi.org/10.3390/su12062237

Chapman, P., Clinton J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step Data Mining Guide*. SPSS Inc. (USA).

Chen, Y. (2021). An Analytical process of spatial autocorrelation functions based on Moran's Index. *PLOS ONE*, *16*, 1-27. https://doi.org/10.1371/journal.pone.0249589

Chong, M., Abraham, A. & Paprzycki, M. (2005). Traffic Accident Analysis Using Machine Learning Paradigms. *Informatica*, *29*, 89-98.

DGT - Direção Geral do Território. (2022, June 30). *Quem Somos*. https://www.dgterritorio.gov.pt/dgt/quem-somos

Dogru, N. & Subasi, A. (2012). Traffic Accident Detection By Using Machine Learning Methods. *Information Systems and Sustainability*, *2*, 468- 474.

Efron B. & Tibshirani R. (1986). Booststrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science*, *1*, 54 -77. DOI: 10.4236/ojs.2015.56052

Fu, X., Meng, H., Wang, X., Yang, H. & Wang, J. (2022). A hybrid neural network for driving behavior risk prediction based on distracted driving behavior data. *PLoS ONE, 17(1)*, 1-17. https://doi.org/10.1371/journal.pone.0263030

Gan, J., Li, L., Zhang, D., Yi, Z. & Xiang, Q. (2020). An Alternative Method for Traffic Accident Severity prediction: Using Deep Forest Algorithm. *Journal of Advanced Transportation*, *2020*,1 – 13. https://doi.org/10.1155/2020/1257627

García, J. A. S., Ortis, A. F. A & García, A. J. S. (2020). Análisis Espacial Para Interpretar La Relación Entre Economía Y Territorio En Xalapa Veracruz A Través Del Geoprocesamiento De Autocorrelación Espacial Índice De Moran Como Diseño Metodológico En La Formación Del Arquitecto. *DAYA. Diseño, Arte y Arquitectura*, *8*, 73-98.

GEODA. (2022, August 15). *Introducing GeoDa 1.20*. https://geodacenter.github.io/

George, S. & Santra, A. K. (2020, July, 22). Traffic Prediction Using Multifaceted Techniques: A Survey. *Wireless Personal Communications*, *115*, 1047–1106. https://doi.org/10.1007/s11277-020-07612-8

Guerreiro, T. M. (2008). *Análise da Sinistralidade Rodoviária em Portugal. Estudo de duas vias: EN6 e A5* (Tese de mestrado). Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa.

Hasheminejad, S. H., Zahedi, M. & Hasheminejad, S. M. H. (2017). A hybrid clustering and classification approach for predicting crash injury severity on rural roads. *International Journal of Injury Control and Safety Promotion*, *25* (1), 85-101. DOI: 10.1080/17457300.2017.1341933

Here. (2022, June 07). *About HERE Technologies*. https://www.here.com/company/about-us

Hong, J. W. (2020). Why Is Artificial Intelligence Blamed More? Analysis of Faulting Artificial Intelligence for Self-Driving Car Accidents in Experimental Settings. *International Journal of Human-Computer Interaction*, *36* (18), 1768-1774. https://doi.org/10.1080/10447318.2020.1785693

IBM (2021). IBM SPSS Modeler 18.3: Algorithms Guide.

IPMA. (2022, September 30). https://www.ipma.pt/pt/index.html

Khatri, S., Vachhani, H., Shah, S., Bhatia, J. Chaturvedi, M., Tanwar, S. & Kumar, N. (2020). Machine learning models and techniques for VANET based traffic management: Implementation issues and challenges. *Peer-to-Peer Networking and Applications*, *14*, 1778–1805. DOI:10.1007/s12083-020-00993-4

MARKTEST (2022, June 07), https://www.marktest.com/wap/

Oetomo, H., W., Lestariningsih, M. & Susanti. (2017). Spatial Analysis of Newspaper Sales in East Surabaya Traffic Lights Using Moran Index. *International Journal of Business and Administrative Studies*, *3* (5), 166-174. DOI: 10.20469/ijbas.3.10002-5

Pereira, P.M.S. (2016). *A sinistralidade rodoviária em ambiente urbano: a cidade de Lisboa como objeto de estudo.* (Tese de mestrado). Instituto Superior de Ciências Policiais e Segurança Interna, Lisboa.

PORDATA. (2022, June 30). *A PORDATA*. https://www.pordata.pt/sobre+a+pordata

QGIS. (2022, July 15). https://qgis.org/en/site/

Rezaein, A., Shokohyar, S. & Zolfaghari, S. (2016). Clustering and Classification of Road Accidents in Iran Using Data Mining Techniques. *International Journal of Business and Information*, *11* (3), 365-383.

Shweta, Yadav, J., Batra, K. & Goel, K. (2021). A Framework for Analyzing Roads Accidents Using Machine Learning Paradigms. *Journal of Physics: Conference Series*, *1950*, 1-8. DOI 10.1088/1742-6596/1950/1/012072

Torrão, G., Coelho, M. & Rouphail, N. (2010). Effect Of Vehicle Characteristics On Crash Severity: Portuguese Experience. *WCTR*, *12*, 1-17. DOI:10.1136/injuryprev-2012-040590u.41

Wang, J. & Chen, Q. (2021). A traffic prediction model based on multiple factors. *The Journal of Supercomputing*, *77*, 2928-2960.

Wehrens R., Putter, H. & Buydens, L. M. C. (2000). The bootstrap: a tutorial. *Chemometrics and Intelligent Laboratory Systems*, *11*, 35-52.

World Health Organization. (2022, June 20). *Road Traffic Injuries*. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

Wirth, R. & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Ming.

Xu, P. & Meng, X. (2019) A novel ensemble learning method for crash prediction using road geometric alignments and traffic data. *Journal of Transportation Safety & Security*,*12*, 1128-1146. https://doi.org/10.1080/19439962.2019.1579288

Annex
Annex A – Variables Independent and Dependent

Univariate Statistics			
	N	Mean	Std. Deviation
alt_min	278	85.0288	116.32886
altitude_med	278	268.6897	213.86185
alt_max	278	648.2158	411.34681
amplitude_media_temperatura	278	22.5831	3.18798
seg_fun1_p_Nseg	278	3.8849	7.78301
seg_fun2_p_Nseg	278	16.6071	24.37151
seg_fun3_p_Nseg	278	70.7008	42.90255
seg_fun4_p_Nseg	278	130.1664	45.72443
seg_fun5_p_Nseg	278	778.6408	51.50521
seg_vel2e3_p_Nseg	278	9.5211	10.86058
seg_vel4e5_p_Nseg	278	180.3559	77.31504
seg_vel6_p_Nseg	278	506.2633	126.35841
seg_vel7e8_p_Nseg	278	303.8596	101.17164
seg_vel2e3_tunel_p_Nseg	278	.0324	.14822
seg_vel4e5_tunel_p_Nseg	278	.0275	.31664
seg_vel6_tunel_p_Nseg	278	.0290	.21874
seg_vel7e8_tunel_p_Nseg	278	.0021	.01949
seg_vel2e3_urban_p_Nseg	278	.9983	3.08494
seg_vel4e5_urban_p_Nseg	278	9.8042	8.86249
seg_vel6_urban_p_Nseg	278	481.5496	126.95324
seg_vel7e8_urban_p_Nseg	278	149.9634	87.17880
n_ped_e_rodo_p_Nseg	278	945.1920	39.45014
auto_moto_p_Nseg	278	1940.5929	62.90721
urban_rodo_p_Nseg	278	660.4741	151.90154
tunel_p_Nseg	278	.0910	.42868
intersecoes_p_Nseg	278	5.1426	3.62621
fun1_p_comp	278	12.3900	24.52306
fun2_p_comp	278	20.7701	24.28690
fun3_p_comp	278	48.1596	28.86601
fun4_p_comp	278	104.2378	40.15510
fun5_p_comp	278	814.4425	49.66324
vel2e3_p_comp	278	27.4096	30.07874
vel4e5_p_comp	278	249.1341	100.61043
vel6_p_comp	278	287.1519	160.94112

vel7e8_p_comp	278	436.3044	150.91219
auto_moto_p_comp	278	1966.7613	49.87992
tunel_p_comp	278	.1562	.71451
intersecoes_p_comp	278	1.1528	1.16323
vel2e3_tunel_p_comp	278	.0887	.52947
vel4e5_tunel_p_comp	278	.0327	.37481
vel6_tunel_p_comp	278	.0318	.25828
vel7e8_tunel_p_comp	278	.0029	.02819
vel2e3_urban_p_comp	278	1.7082	4.89777
vel4e5_urban_p_comp	278	5.9338	6.11842
vel6_urban_p_comp	278	267.5362	158.03794
vel7e8_urban_p_comp	278	91.6023	64.32470
urban_rodo_p_comp	278	402.7826	218.13792
seg_vel2e3_urban_p_Nurban	278	.1255	.35002
seg_vel4e5_urban_p_Nurban	278	1.6224	1.55303
seg_vel6_urban_p_Nurban	278	75.3403	11.32737
seg_vel7e8_urban_p_Nurban	278	22.9118	11.29022
vel2e3_urban_p_compurban	277	.2785	.66076
vel4e5_urban_p_compurban	277	1.7746	1.52749
vel6_urban_p_compurban	277	73.2684	10.97050
vel7e8_urban_p_compurban	277	24.6785	10.95065
População2018	278	35179.2302	57227.79425

Source: Here, PORDATA, DGT and IPMA

	N	Mean	Std. Deviation
ac_vit_auto_p_comp	278	.0074	.01916
ac_vit_mort_auto_p_comp	278	.0002	.00043
ac_vit_mort_nac_p_comp	278	.0007	.00108
ac_vit_mort_p_comp	278	.0021	.00244
ac_vit_nac_p_comp	278	.0247	.04384
tot_ac_p_comp	278	.1217	.19573
fer_grav_p_pop	278	.3246	.31060
fer_lig_p_pop	278	4.1219	1.37480
vit_auto_p_pop	278	.2846	.53514
vit_mort_p_pop	278	.1121	.15709
vit_nac_p_pop	278	1.4751	1.05665
tot_vit_p_pop	278	4.5586	1.49434

Source: INE

Annex B - Rotated Component Matrix – Independent Variables

Rotated Component Matrix[a]													
	Component												
	1	2	3	4	5	6	7	8	9	10	11	12	13
vel6_p_comp	.909	-.056	.281	-.005	.151	.003	.087	.089	-.039	.030	.007	-.079	.010
urban_rodo_p_Nseg	.908	.254	.171	-.054	.092	.036	.048	.077	-.082	.051	-.003	-.028	.008
vel6_urban_p_comp	.907	-.053	.291	.002	.164	-.002	.081	.074	-.025	.026	-.007	-.086	.014
seg_vel6_urban_p_Nseg	.892	-.336	.150	-.113	.093	.012	.048	.045	-.006	.013	-.050	-.017	-.010
seg_vel6_p_Nseg	.891	-.349	.137	-.123	.079	.020	.053	.069	-.026	.016	-.025	.003	-.014
urban_rodo_p_comp	.886	.184	.295	.032	.181	.083	.077	.073	-.015	.059	.046	-.083	.005
vel7e8_p_comp	-.815	.136	-.057	-.083	.105	.012	-.024	-.264	.220	-.032	-.244	-.036	-.194
seg_vel4e5_p_Nseg	-.672	-.303	-.258	.271	-.210	-.108	-.068	.162	-.042	-.061	.214	.189	.181
seg_vel7e8_urban_p_Nurban	-.104	.972	-.050	-.015	-.053	-.049	-.020	.017	-.121	.021	.007	-.037	.012
vel7e8_urban_p_compurban	-.129	.967	-.068	-.027	-.037	-.013	-.023	-.015	-.033	.009	-.032	-.023	-.020
seg_vel6_urban_p_Nurban	.126	-.964	.024	-.114	.054	.042	.020	-.022	.118	-.021	-.013	.033	-.015
vel6_urban_p_compurban	.144	-.962	.020	-.101	.027	.002	.015	.007	.043	-.014	.018	.024	.015
seg_vel7e8_urban_p_Nseg	.233	.941	-.019	-.030	-.004	-.059	-.023	.034	-.124	.028	-.012	-.028	.017
vel7e8_urban_p_comp	.559	.744	.085	.032	.121	-.043	-.021	.003	.007	.035	-.046	-.063	-.010
seg_vel7e8_p_Nseg	-.620	.670	-.020	-.045	.062	-.016	-.014	-.214	.063	.007	-.167	-.146	-.122
por100	.269	-.057	.911	-.016	.052	.077	.035	.031	-.063	.088	.095	-.009	.164
seg_vel2e3_urban_p_Nseg	.271	-.049	.907	-.006	.069	.048	.033	.020	-.076	.073	.083	-.004	.168
vel2e3_urban_p_comp	.323	-.036	.892	.006	.114	.046	.053	.000	-.085	.062	.085	-.011	.138
vel2e3_urban_p_compurban	.301	-.043	.874	-.022	.061	.129	.045	.011	-.064	.088	.121	-.024	.093
vel4e5_p_comp	-.299	-.125	-.427	.141	-.396	-.227	-.094	.237	-.296	-.041	.246	.188	.296
seg_vel4e5_urban_p_Nseg	.055	.067	-.024	.972	-.006	.052	.013	.031	-.022	-.004	-.001	.019	.018
seg_vel4e5_urban_p_Nurban	-.224	-.020	-.015	.945	-.019	.032	-.010	.033	.035	-.020	.021	.032	-.018
vel4e5_urban_p_compurban	-.239	-.005	-.033	.927	.046	.026	.035	.050	-.041	-.002	.050	-.001	-.006
vel4e5_urban_p_comp	.373	.147	.036	.783	.243	.007	.147	.055	-.155	.046	.003	-.011	.070
vel6_tunel_p_comp	.127	.060	.054	.088	.855	-.026	.040	-.017	-.255	.064	-.055	.110	.111
seg_vel6_tunel_p_Nseg	.125	.061	.059	.076	.850	-.017	.039	-.001	-.242	.068	-.052	.143	.102
intersecoes_p_comp	.369	-.101	.142	.028	.753	.032	.177	.102	-.182	-.040	.175	-.013	.098
intersecoes_p_Nseg	.046	-.241	.053	-.007	.643	.051	.091	.073	-.090	-.116	.299	.007	.091
fun1_p_comp	-.019	-.073	-.036	.097	.007	.929	-.019	-.017	.046	-.050	.004	-.073	.040
por milhares de segmentos	-.097	-.095	.046	.062	.020	.926	-.005	-.045	-.007	.012	-.053	-.066	.078
vel2e3_p_comp	.267	.035	.230	-.034	-.003	.731	-.028	.057	.097	.145	.388	-.030	-.076
seg_vel2e3_p_Nseg	.192	-.025	.439	-.075	.001	.683	-.007	.045	.016	.184	.328	-.024	.001
seg_vel4e5_tunel_p_Nseg	.089	-.051	.033	.038	.022	-.022	.987	.046	-.037	-.009	.079	.013	-.015

vel4e5_tunel_p_comp	.092	-.054	.025	.043	.030	-.024	.986	.040	-.044	-.011	.077	.012	-.022
tunel_p_Nseg	.144	-.006	.141	.063	.432	.008	.750	.021	-.168	.329	.036	.080	.123
seg_fun4_p_Nseg	.161	-.040	.065	.074	.069	-.036	.027	.898	.092	-.008	-.120	-.243	-.044
fun4_p_comp	.278	.128	-.085	-.001	-.081	-.160	.006	.854	-.163	-.029	-.045	-.158	.030
seg_fun5_p_Nseg	.158	.213	-.099	-.147	-.153	-.096	-.083	-.702	-.146	.017	-.188	-.439	.031
fun5_p_comp	-.314	-.063	-.007	-.028	-.052	-.257	-.085	-.701	.103	-.053	-.323	-.356	-.001
auto_moto_p_Nseg	-.009	-.227	.005	-.017	-.205	.135	-.042	-.006	.889	-.028	.050	.019	-.030
n_ped_e_rodo_p_Nseg	-.117	-.119	-.175	-.041	-.191	-.112	-.101	-.043	.800	-.104	-.108	.010	-.046
auto_moto_p_comp	-.133	-.063	-.078	-.073	-.341	.089	-.073	.029	.762	-.030	.082	.056	.040
vel2e3_tunel_p_comp	.035	.063	.020	-.008	-.026	.025	-.002	.005	-.025	.948	.084	-.008	-.014
seg_vel2e3_tunel_p_Nseg	.040	.006	.220	-.013	-.063	.089	.002	-.031	-.050	.860	.016	-.005	.122
tunel_p_comp	.123	.041	.058	.050	.319	-.002	.531	.019	-.136	.723	.081	.041	.048
fun2_p_comp	.176	.038	.257	-.020	.072	.069	.150	.035	.034	.188	.842	-.038	-.070
seg_fun2_p_Nseg	-.143	-.110	.080	.091	.104	.166	.068	.008	-.006	-.014	.831	-.149	-.001
fun3_p_comp	.025	-.045	-.061	-.007	.136	-.092	.028	.017	-.016	.011	-.093	.921	-.011
seg_fun3_p_Nseg	-.261	-.134	-.004	.035	.048	-.105	.033	-.110	.082	-.006	-.108	.882	-.003
seg_vel7e8_tunel_p_Nseg	.011	-.023	.238	-.003	.078	.043	.019	-.046	.011	.060	-.028	-.032	.882
vel7e8_tunel_p_comp	.076	.012	.266	.042	.347	.051	.015	.018	-.055	.088	-.043	.013	.762

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.[a]

a. Rotation converged in 8 iterations.

Annex C – Correlation Table

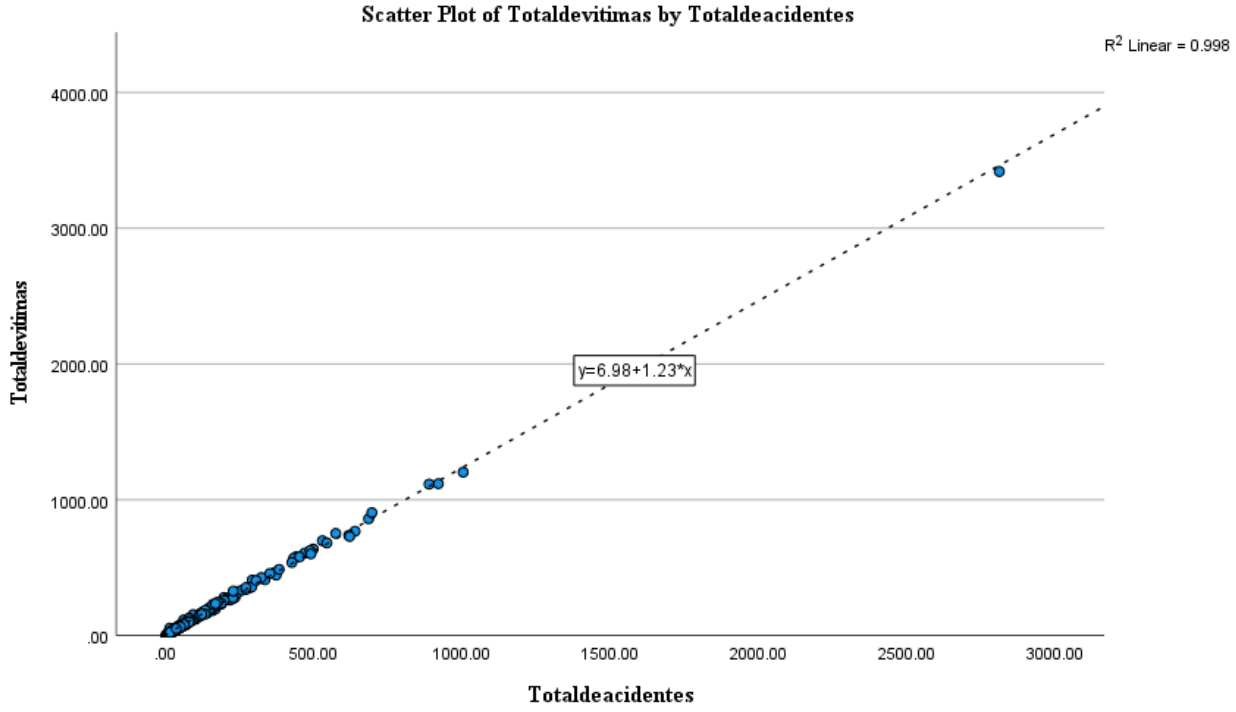
Correlations

		Totaldevitimas	Totaldeacidentes
Totaldevitimas	Pearson Correlation	1	.999**
	Sig. (2-tailed)		.000
	N	278	278
Totaldeacidentes	Pearson Correlation	.999**	1
	Sig. (2-tailed)	.000	
	N	278	278

**. Correlation is significant at the 0.01 level (2-tailed).

Source: INE

Annex D – Scatter Plot of the victims by accidents



Source: INE

Annex E – Causality Test: accidents and victims, absolute values

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.999[a]	.998	.998	12.02666
a. Predictors: (Constant), Totaldeacidentes_sum				

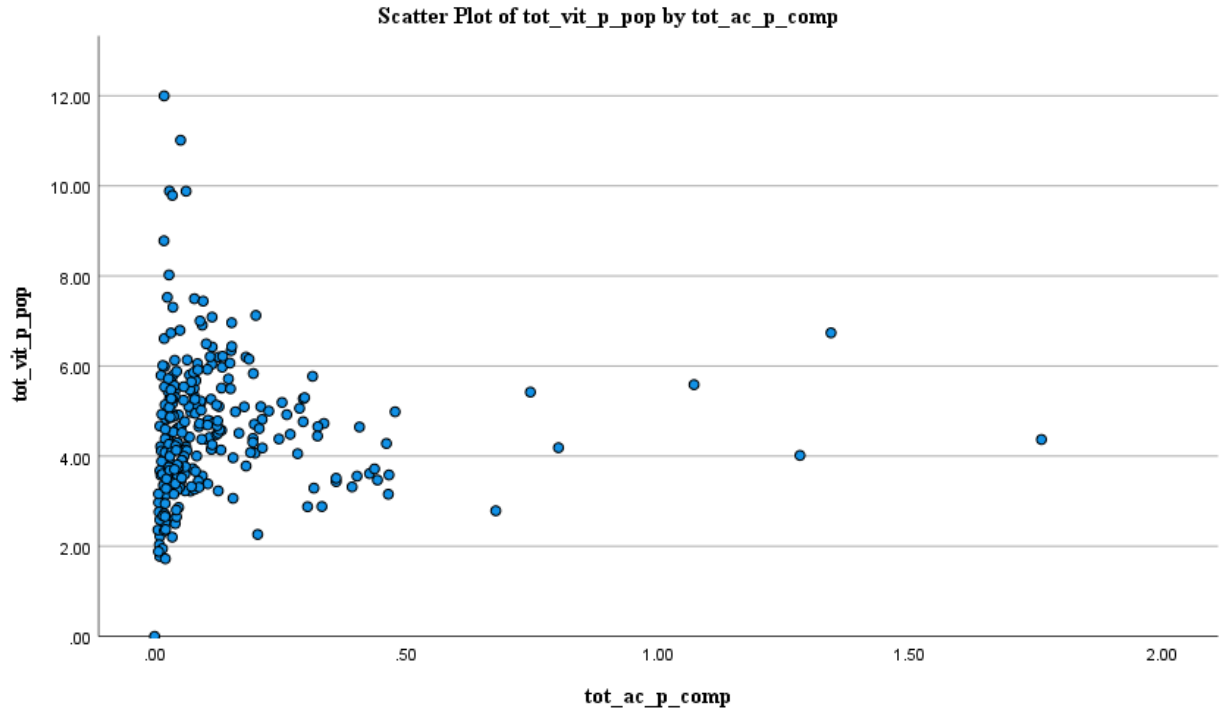
Coefficients[a]						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6.978	.820		8.507	.000
	Totaldeacidentes_sum	1.229	.003	.999	387.455	.000
a. Dependent Variable: Totaldevitimas_sum						

Annex F – Correlation Tables: tot_ac_p_comp and tot_vit_p_pop

Correlations			
		tot_vit_p_pop	tot_ac_p_comp
tot_vit_p_pop	Pearson Correlation	1	.036
	Sig. (2-tailed)		.554
	N	278	278
tot_ac_p_comp	Pearson Correlation	.036	1
	Sig. (2-tailed)	.554	
	N	278	278
**. Correlation is significant at the 0.01 level (2-tailed).			

Source: INE

Annex G – Scatter Plot of victims per 1000 inhabitants by accidents per 1000 meters



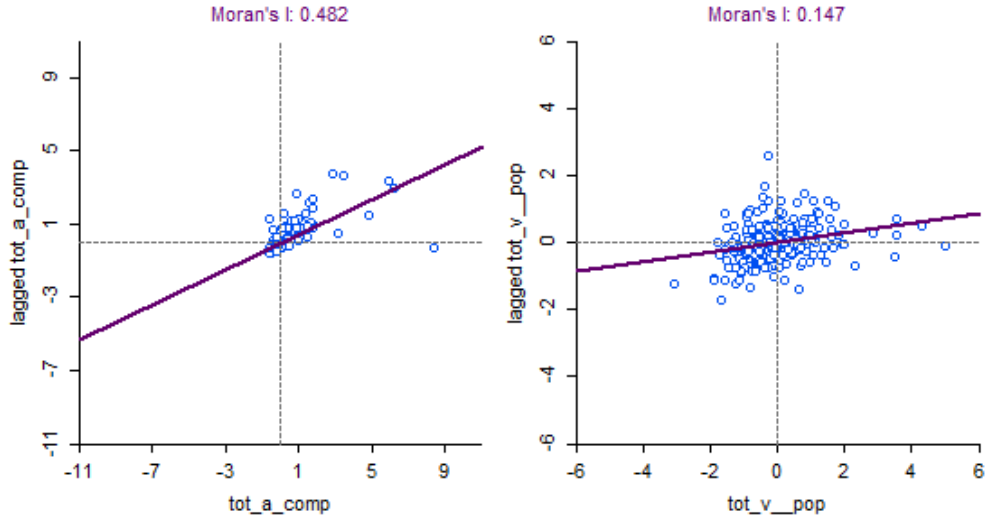
Annex G – Causality Test: tot_ac_p_comp and tot_vit_p_pop

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.148[a]	.022	.008	1.488561469504802
a. Predictors: (Constant), A3, w_acid, tot_a_comp, A2				

ANOVA[a]						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	13.634	4	3.409	1.538	.191[b]
	Residual	604.918	273	2.216		
	Total	618.552	277			
a. Dependent Variable: tot_v__pop						
b. Predictors: (Constant), A3, w_acid, tot_a_comp, A2						

Coefficients[a]						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4.455	.150		29.609	.000
	tot_a_comp	4.652	2.067	.609	2.250	.025
	w_acid	-2.525	1.294	-.226	-1.951	.052
	A2	-4.511	3.791	-.761	-1.190	.235
	A3	1.310	1.853	.339	.707	.480
a. Dependent Variable: tot_v__pop						

Annex H - Moran's Index



Annex I – Linear Regression with Spatial Lag: tot_a_comp

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : municipalities
Spatial Weight      : municipalities
Dependent Variable  :  tot_a_comp  Number of Observations:  278
Mean dependent var  :    0.121739  Number of Variables   :    8
S.D. dependent var  :    0.195374  Degrees of Freedom    :  270
Lag coeff.   (Rho)  :    0.329415

R-squared           :    0.702745  Log likelihood        :      224.95
Sq. Correlation     : -            Akaike info criterion :    -433.901
Sigma-square        :   0.0113466  Schwarz criterion     :     -404.88
S.E of regression   :     0.10652

------------------------------------------------------------------------------
       Variable       Coefficient     Std.Error       z-value    Probability
------------------------------------------------------------------------------
      W_tot_a_comp      0.329415      0.0620842        5.30593     0.00000
          CONSTANT    -0.0667859      0.0145295       -4.59658     0.00000
         vel6_comp   0.000256695    5.79686e-05        4.42816     0.00001
           vel6_tu      0.148413      0.0379175        3.91412     0.00009
        inter_comp     0.0245206     0.00901599        2.71968     0.00653
        fun1__comp    0.00211286    0.000267128        7.90955     0.00000
          vel4e5_t      0.111276       0.018234        6.10269     0.00000
        Popul_2018   4.04026e-07    1.82029e-07        2.21956     0.02645
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                    DF      VALUE        PROB
Breusch-Pagan test                      6     10291.1672     0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : municipalities
TEST                                   DF      VALUE        PROB
Likelihood Ratio Test                  1        17.8992     0.00002
============================== END OF REPORT ================================
```

Annex J - Linear Regression with Spatial Lag: tot_v_pop

```
>>10/11/22 22:34:28
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : municipalities
Spatial Weight      : municipalities
Dependent Variable  :  tot_v__pop  Number of Observations:  278
Mean dependent var  :      4.5586  Number of Variables   :    6
S.D. dependent var  :     1.49165  Degrees of Freedom    :  272
Lag coeff.   (Rho)  :    0.174205

R-squared           :    0.141594  Log likelihood        :   -485.244
Sq. Correlation     : -            Akaike info criterion :    982.488
Sigma-square        :     1.90996  Schwarz criterion     :    1004.25
S.E of regression   :     1.38201

------------------------------------------------------------------------------
       Variable       Coefficient     Std.Error       z-value    Probability
------------------------------------------------------------------------------
      W_tot_v__pop      0.174205      0.0856306        2.03437     0.04191
          CONSTANT       6.35173       0.916945        6.92706     0.00000
         Popul_2018  -5.87319e-06    1.94912e-06       -3.01325     0.00258
         funl__comp     0.0104257     0.00341678        3.05132     0.00228
            vel6_tu       1.17124       0.393908        2.97339     0.00295
         ampli_tura     -0.112784      0.0303984       -3.71019     0.00021
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                DF      VALUE        PROB
Breusch-Pagan test                  4        27.3453     0.00002

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : municipalities
TEST                                DF      VALUE        PROB
Likelihood Ratio Test               1         4.0790     0.04342
```

Annex K – Predictor Importance: Accidents Models

Model	Variable	Predictor Importance
Neural Network	w_acid	0.31
	vel6_tu	0.14
	vel4e5_tu	0.14
	Inter_comp	0.06
	popu_2018	0.04
	fun1_comp	0.11
	vel6_comp	0.02
Regression	w_acid	0.39
	vel4e5_tu	0.18
	Inter_comp	0.13
	vel6_tu	0.12
	fun1_comp	0.07
	Popu_2018	0.05
	vel6_comp	0.05
Random Tree	popu_2018	0.12
	vel6_comp	0.01

	inter_comp	0.10
	fun1_comp	0.20
	w_acid	0.22
	vel6_tu	0.13
	vel4e5_tu	0.22
Cart Boosting	vel4e5_tu	0.15
	vel6_tu	0.15
	inter_comp	0.15
	w_acid	0.15
	vel6_comp	0.15
	popu_2018	0.15
	fun1_comp	0.07
Cart Bagging	vel4e5_tu	0.15
	vel6_tu	0.15
	inter_comp	0.15
	w_acid	0.15
	vel6_comp	0.15
	popu_2018	0.15
	fun1_comp	0.07
LSVM	w_acid	0.24
	inter_comp	0.2
	popu_2018	0.18
	vel4e5_tu	0.12
	vel6_comp	0.09
	vel6_tu	0.09
	fun1_comp	0.07
SVM	w_acid	0.51
	vel6_tu	0.22
	vel4e5_tu	0.18
	fun1_comp	0.08
	vel6_comp	0.01
	inter_comp	0
	popu_2018	0

Annex L – Predictor Importance: Severity Models

Models	Variables	Predictor Importance
Neural Network	popu_2018	0.29
	fun1_comp	0.16
	vel6_tu	0.17
	w_vit	0.21
	amplitude térmica	0.17
Regression	fun1_comp	0.24
	amplitude térmica	0.27
	vel6_tu	0.07
	w_vit	0.35
	popu_2018	0.05
Random Tree	popu_2018	0.29
	fun1_comp	0.14
	vel6_tu	0.25
	amplitude térmica	0.18
	w_vit	0.14
Cart	w_vit	0.49
	amplitude térmica	0.31
	fun1_comp	0.17
	popu_2018	0.02
	vel6_tu	0.02
Cart Boosting	w_vit	0.25
	fun1_como	0.25
	amplitude térmica	0.25
	popu_2018	0.25
	vel6_tu	0.00
Cart Bagging	popu_2018	0.23
	fun1_comp	0.23
	amplitude térmica	0.23
	w_vit	0.23
	vel6_tu	0.09
Chaid Boosting	amplitude térmica	0.26
	w_vit	0.26
	popu_2018	0.23
	vel6_tu	0.00
	fun1_comp	0.26
Chaid Bagging	vel6_tu	0.11
	amplitude térmica	0.29
	popu_2018	0.03

	fun1_comp	0.29
	w_vit	0.29
LSVM	popu_2018	0.03
	fun1_comp	0.19
	w_vit	0.38
	vel6_tu	0.14
	amplitude térmica	0.26