Accidents' point of incidence in Lisbon

Beatriz Macedo[1], Frederico Ferreira[2], Ricardo Martins[3]

[1] Student number: 20201719, e-mail: m20201719@novaims.unl.pt
[2] Student number: 20201723, e-mail: m20201723@novaims.unl.pt
[3] Student number: 20201443, e-mail: m20201443@novaims.unl.pt

Abstract: The main objective of this project was identifying the geographical areas with higher prevalence of accidents in 2019. Using several datasets provided by the LxDataLab and geographic data from the ArcGIS Lisboa, it was possible to identify the parishes with higher frequency of accidents. Further, a linear regression model was used to compare the real number of accidents registered for each parish with the predicted value, therefore identifying the ones with a misadjusted number of accidents for their endogenous characteristics. It was possible to conclude that Avenidas Novas and Alvalade register a higher frequency of accidents, which may be explained by characteristics of these parishes and therefore it cannot be implied that these parishes are more dangerous than others.

Keywords: Traffic accidents; Accident risk; Geodata

I. Introduction

Urban mobility plays an important role in addressing urban livability. Traffic accidents occur every day and besides the monetary and human damage that often result from them, they also have an important impact in the urban mobility of several other people, therefore affecting the livability of the cities. Accidents often occur because of human error, however, external factors such as the condition of the road, the presence or absence of road signaling and the meteorological conditions, among others, can affect or increase their occurrence since they can produce changes in perception of the drivers [1].

As such, the Lisbon Municipality launched a challenge that aims to identify the points with the highest incidence of road accidents and their correlation with other factors, namely signaling, orientation and inclination of roads.

In this sense, this report focuses on identifying the geographical areas with higher prevalence of accidents using geo-spatial visualization technics (GeoPandas and ArcGIS). Also, regression techniques were used to control endogenous factors of each parish, therefore allowing to understand the real risk of accident in each of them.

II. Data

a. Description

The data used in this project was provided by the Lisbon Municipality and is comprised by the following datasets:

- Road Accident Occurrences (RSB);
- Accident data from the National Road Safety Agency (ANSR);
- Slope of roads;
- Horizontal and vertical signage;
- Traffic lights;

- Metro stations;
- Altimetry; and,
- Traffic data (Waze).

The data includes information about the accidents that occurred in the Lisbon Municipality during 2019. For each accident there is information about the date and time of its occurrence, the location of the accident (GPS coordinates), parish of occurrence, nature of the accident (runover, collision, or overturn), road conditions (straight lines or curves, leveled or inclined roads, sideroad and pavement conditions, obstacles, presence of vertical and lighting signals and intersections) and finally the meteorological and lighting conditions and adherence to the road. Information about the vehicles and drivers involved in the accident is also available (type and condition of vehicle, age, gender, health outcomes and behavior). Information about other passengers is also available (gender, age and health outcomes). Finally, for the runover accidents, information about the pawns that were involved is also presented (age, gender, health outcomes and actions that the pawns were performing when the accident occurred).

Besides the data provided by the Lisbon Municipality and to complement the spatial representation and analysis of the accidents in Lisbon, geographical data regarding the limits of the several parishes[1] was obtained from the Lisbon Municipality website [2].

Further to the above, the following information was gathered for each parish in the Lisbon Municipality from the Instituto Nacional de Estatística (INE)[2]:

- Total population per parish;
- Active population;
- Number of unemployed;
- Preferred mean of transportation when commuting.

Additionally, also from INE, the mean housing price per square meter for 2019 was obtained.

b. Extraction

Most data were provided by the Lisbon Municipality, under the LxDataLab program. Additional geographical data was extracted from AcrGIS Lisboa. The data from INE was obtained from their website.

c. Transformation

Pre-processing

The ANSR dataset is composed of an excel file with 4 spreadsheets, with the first one containing the accidents, the second one containing the vehicle information and the third and fourth containing the passenger and pawn information, respectively. The common variable among the four spreadsheets is the “IdAcidente”, which was used as index to join the information between them. In the vehicle, passenger, and pawn datasets each accident can have more than one record.

[1] The data was last updated in 19 of March 2021.

[2] The most recent data available is from the national statistics institute (Censos 2011).

This dataset did not contain duplicates and only the “Dia da Semana” feature required the removal of spaces (trim).

The INE data is composed of two excel files containing data per parish (prior to 2013 reorganization). The main pre-processing for this data was the conversion of the 53 parishes (2011) to the 23 (2013 onwards). In this process, the conversion table required text splitting to single out the parish name.

The RSB dataset was not considered for analysis as it does not provide additional information, when compared with the ANSR file. Additionally, the traffic data from Waze was also disregarded since there was no information regarding date and time, which prevented the match between the accident data and the traffic. In addition, the information regarding cross-roads, altimetry and traffic lights was joined to the initial dataset.

Feature engineering

In order to retrieve more information from the dataset, several features were created. Since the vehicle, passenger, and pawn datasets may have more than one record for each “IdAcidente" they were not joined to avoid repeating rows. Instead, new variables were created in the ANSR dataset, which counted the number of vehicles, passengers and pawns in each accident.

In addition, the “datahora” column was converted to datetime enabling the creation of a new feature by binning the part of day (“morning”, “afternoon”, “evening” and “night”).

With the variables “Longitude GPS” and “Latitude GPS” it was possible to create a GeoPandas dataframe which added a new feature to the dataset named “geometry”. Using this feature, it was possible to join the spatial information of the accident with other geodatasets. For example, to match each accident to the parish polygon the intersection between the accidents location and the parishes was calculated.

As for missing values, since the altimetry feature had several missing observations, it was dropped before further analysis. Also, there is GPS information missing for 945 accidents. Therefore, when geospatial information for accidents was required, these records were not considered.

To complete the parish characterization, a Pandas dataframe was created containing the demographic information from INE, where the following features were added, per parish:

- Number of accidents,
- Geographic polygon,
- Number of metro stations,
- Parish area in m^2,
- Median housing price/m^2,
- Total population,
- Active population,
- Unemployed population,
- Number of people commuting with each type of vehicle,
- Number of cross-roads,

- Number of traffic lights.

Having the geographic polygon of each parish, the Pandas dataframe mentioned above was converted into a GeoPandas dataframe, which allowed the intersection of the geographical data of the cross-roads and metro stations, enabling the count of occurrence in each parish. The traffic lights variable was dropped since it did not bring additional information when compared to the number of cross-roads.

d. Methodology

Before including the parish characterization variables in the model, the correlation between all the possible pairs of variables was obtained to identify which ones are highly correlated ($\rho > 0.8$). Since the variables have different magnitudes, the dataset was scaled using the StandardScaler. Afterwards, a Recursive Feature Elimination (RFE) was used to statistically select which features should be included in the model (using a linear regression model). Since the optimal number of variables to keep was unknown, a loop was created to obtain the score of the estimators considering each number of features and find the optimal number of variables[3]. After selecting the features, these were used to build a linear regression model and estimate the number of accidents per parish.

III. Results and Discussion

a. Results

In a first instance, the evolution of the number of accidents during the year was evaluated to understand if some periods are more prone to the occurrence of accidents than others (Figure 1). The frequency of accidents remained constant during the entire year.

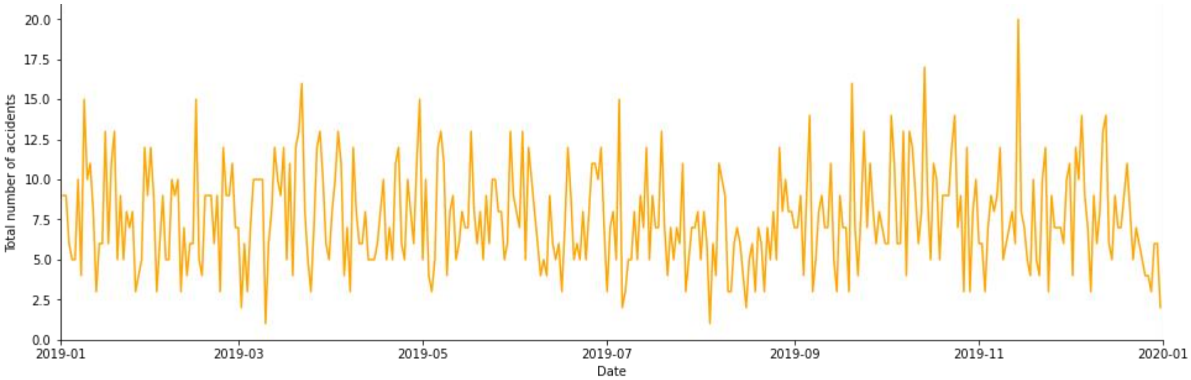


Figure 1 *- Evolution of the number of accidents during 2019.*

Taking into consideration the accident frequency by weekday, accidents occurred mostly from Monday through Friday, while Sunday typically registers less accidents. Fridays registered the higher frequency of accidents, as per Figure 2.

[3] To measure the score python uses the R^2 (by default).

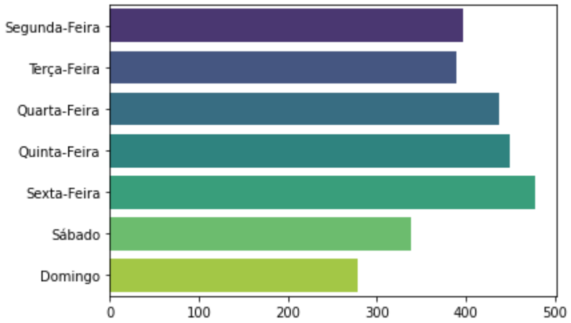


Figure 2 *- Number of accidents by weekday*

Additionally, it was possible to identify the time of day with more accidents, as shown by Figure 3. The frequency is higher during the day, with 8am to 9am registering the most accidents during the morning and 5pm to 7pm registering the highest number in the afternoon. These findings match with the usual times where most individuals commute.

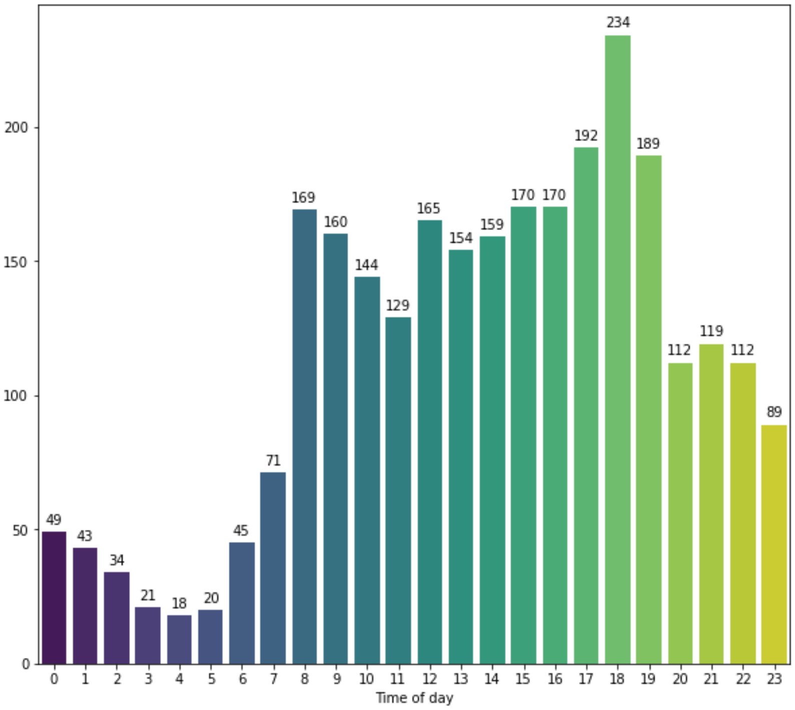


Figure 3 *- Number of accidents by time of day*

Moreover, when joining the weekday and time of accident, it was observed that Wednesday at 6pm registers the highest number of accidents, followed by Monday at 9am. The early hours of Saturday and Sunday also register a considerable number of accidents, probably due to nightlife[4].

The geographical representation of the accidents was plotted in a map with the parishes of Lisbon, to obtain a clear representation of the locations prone to their occurrence (Figure 4).

[4] Analysis is not present here. Further detail is available in the Jupyter notebook.

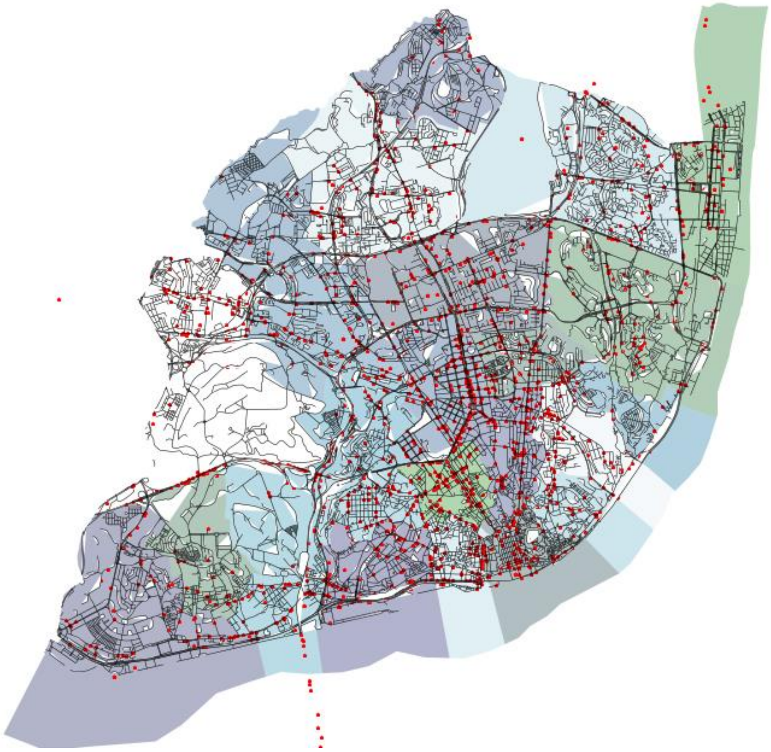


Figure 4 *- Distribution of the accidents in the Parish map of Lisbon (2019)*

Visually, it is possibly to identify that some areas concentrate a higher number of accidents than others. Avenidas Novas and Alvalade registered a much higher number of accidents during 2019 than other parishes such as Beato, Carnide, Santa Clara and Ajuda. To better identify the parishes with higher prevalence a frequency and a heatmap were plotted. The results are shown below in Figure 5.

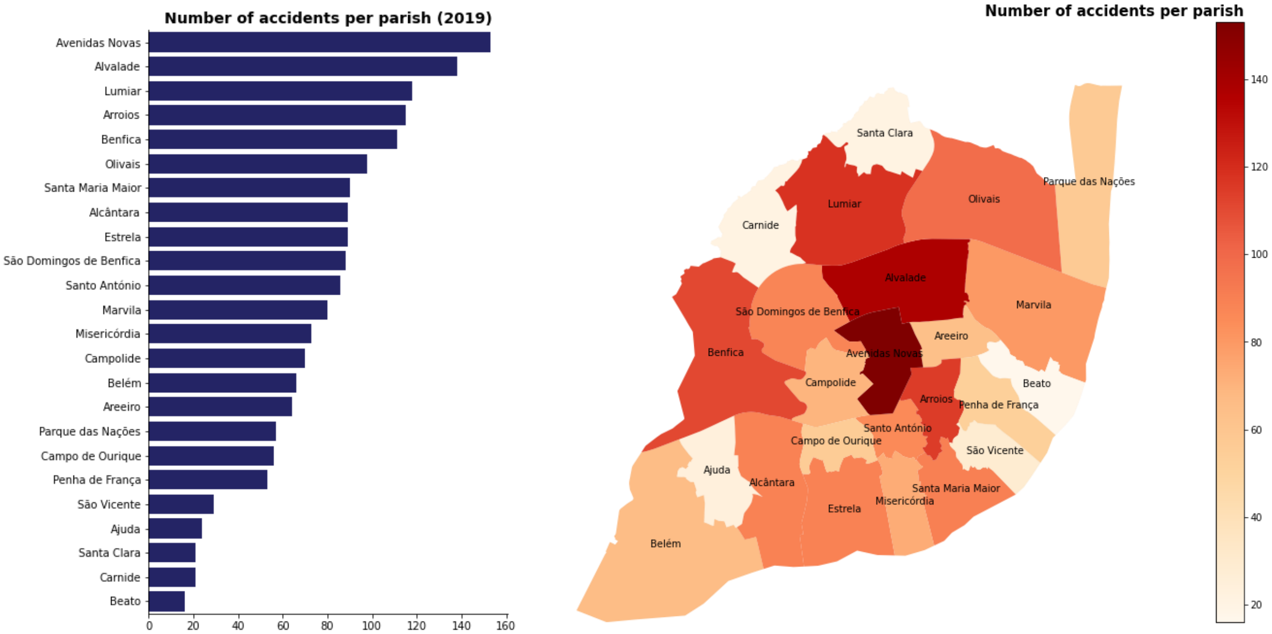


Figure 5 *- Representation of the number of accidents per Parish in bar chart (left) and heatmap (right) in 2019.*

Moreover, the dataset detailed different types of accidents such as collisions (“Colisão”), runovers (“Atropelamento”) and overturns (“Despiste”). During 2019, these types of accidents totaled 1.458, 929 and 553, respectively.

The analysis of the type of accident was also extended by considering the parish where it occurred. In this sense, the number of collisions by parish can be observed in Figure 6, where Avenidas Novas, Lumiar and Benfica are the top 3 parishes, followed by Alvalade and Arroios. On the other hand, Beato, Ajuda and São Vicente register the least collisions. These results follow the same trend as the total accidents per parish.

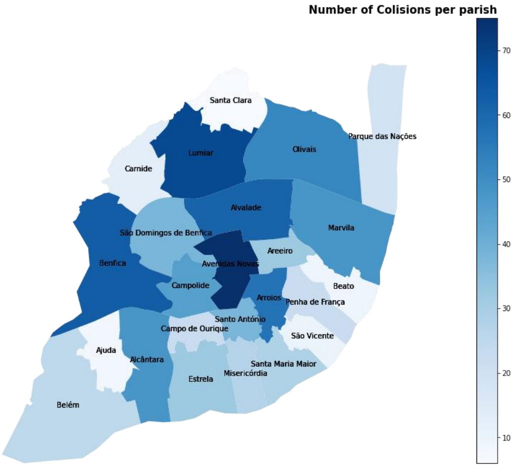


Figure 6 *- Number of collisions by parish*

Regarding runover accidents by parish, besides Avenidas Novas and Alvalade, Santa Maria Maior also stands out in this accident type, as per Figure 7. Conversely, Beato, Carnide and Campolide registered the least runovers.

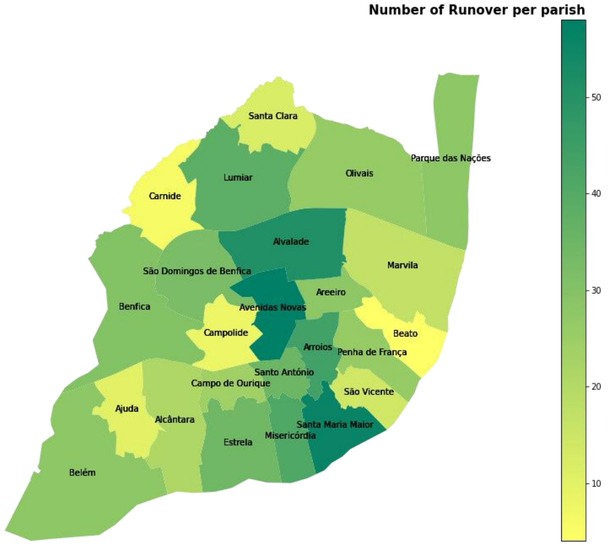


Figure 7 *- Number of runover accidents by parish*

Finally, by plotting the overturn accidents per parish, it is possible to observe that Alvalade and Estrela stand out, according to Figure 8. On the other hand, Beato and Areeiro registered the least occurrences.

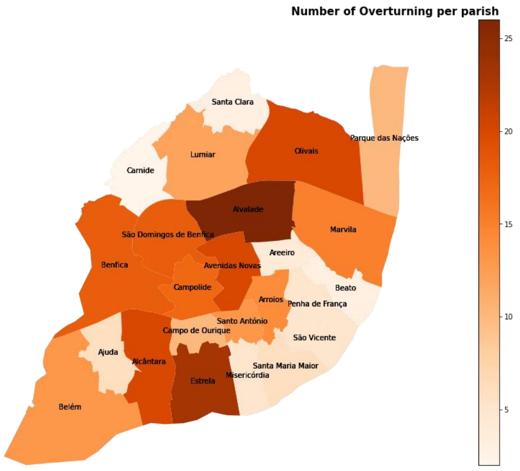


Figure 8 *- Number of overturn accidents by parish*

Further to the above, it is possible to conclude that Alvalade and Avenidas Novas stand out in every category.

Besides the time evolution and geographical occurrence of the accidents, several variables conditioning the occurrence of accidents were analyzed. When considering the properties of the roads, most of the accidents occurred outside highways and in roads in a good or regular state of preservation. As for the floor type, although some accidents occurred in cobblestone roads, most accidents occurred in asphalt. Also, most of the roads were clearly marked (marks separating the different lanes), still, a quarter of accidents occurred in roads poorly/not marked. As per roadside conditions, the majority of accidents occurred in well paved roadsides, however a third of the accidents occurred in roads with roadsides in poor conditions or no roadsides.

When considering road direction, the greater proportion of accidents occurred in straight roads. When considering the inclination, two thirds of the accidents occurred in flat roads.

As per accidents occurring in roads with multiple lanes, accidents occurred more often in the right lane (75% of the accidents). When also considering the presence or absence of road intersections, accidents usually do not occur in intersections (60% of the accidents) and when they do, the more relevant types are junctions and cross-roads. Some accidents also occur in acceleration roads and roundabouts.

Finally, approximately 80% of the accidents occurred in dry conditions and 84% occurred with good weather. Still, a few accidents occurred in wet conditions (approximately 15%). The remaining accidents (not a considerable amount) occurred due to the presence of oil or accumulated water in the road. Finally, 67% of the accidents occurred during the day and 28% during the night, however with good illumination conditions.

Since the majority accidents seem to take place with favorable external conditions, it was not possible to determine which ones influence accidents the most.

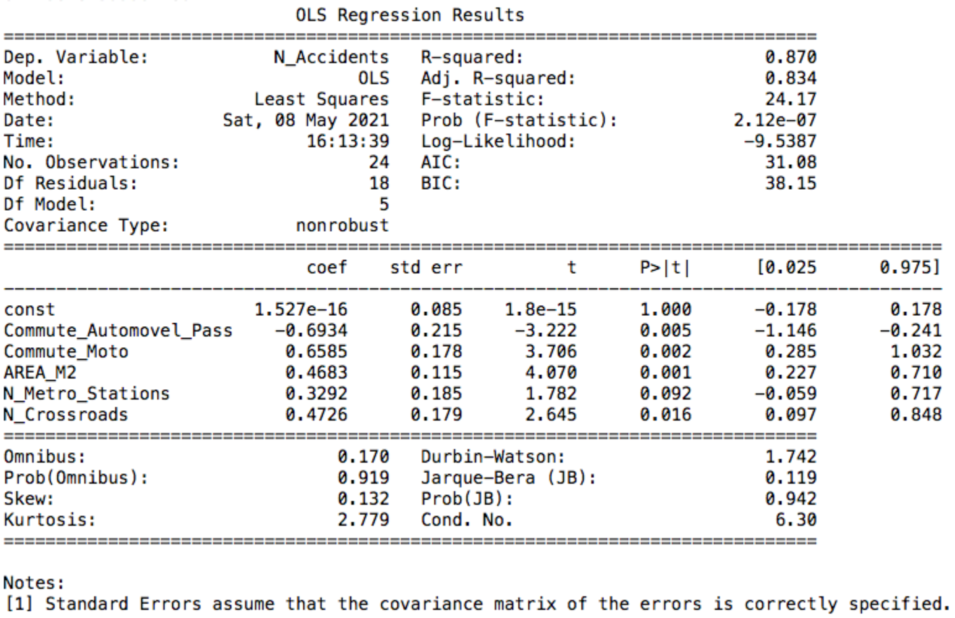
As previously referred, Avenidas Novas and Alvalade are the parishes with the highest number of accidents, however, they are also the ones with the highest population and people commuting. Similarly, Beato and Ajuda are the parishes with fewer accidents and with lower population. This is an indicator that the size of the parish might influence the number of accidents and therefore shows it is necessary to control for endogenous characteristics from the parishes in the model.

All the features described in the parish characterization dataframe were considered in the initial stage of the regression[5].

After analyzing the correlations and selecting the set of pairs with correlation above 0.8, the decision was to drop the variables less correlated with the target (number of accidents). As such, the variables total population, active population, number of unemployed and number of car (driver) and bicycle commutes were not included in the model.

After standardizing the dataset, RFE allowed to conclude that the optimal number of features is 5 and that the variables that should be kept in the model are the parish area (“Area_M2”), number of metro stations (“N_Metro_Stations”), number of crossroads (“N_Crossroads”), number of passenger commutes in cars (“Commute_Automovel_Pass”) and number of commutes in motocycles (“Commute_Moto”).

The regression output of these regressors on the number of accidents is shown below (Table 1).



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            N_Accidents   R-squared:                       0.870
Model:                            OLS   Adj. R-squared:                  0.834
Method:                 Least Squares   F-statistic:                     24.17
Date:                Sat, 08 May 2021   Prob (F-statistic):           2.12e-07
Time:                        16:13:39   Log-Likelihood:                -9.5387
No. Observations:                  24   AIC:                             31.08
Df Residuals:                      18   BIC:                             38.15
Df Model:                           5
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                   1.527e-16      0.085    1.8e-15      1.000      -0.178       0.178
Commute_Automovel_Pass    -0.6934      0.215     -3.222      0.005      -1.146      -0.241
Commute_Moto               0.6585      0.178      3.706      0.002       0.285       1.032
AREA_M2                    0.4683      0.115      4.070      0.001       0.227       0.710
N_Metro_Stations           0.3292      0.185      1.782      0.092      -0.059       0.717
N_Crossroads               0.4726      0.179      2.645      0.016       0.097       0.848
==============================================================================
Omnibus:                        0.170   Durbin-Watson:                   1.742
Prob(Omnibus):                  0.919   Jarque-Bera (JB):                0.119
Skew:                           0.132   Prob(JB):                        0.942
Kurtosis:                       2.779   Cond. No.                         6.30
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

***Table 1** - Regression model output*

By analyzing the regression output it can be concluded that most variables are statistically significant considering a 1% significance level, except for “N_crossroads” and “N_Metro_Stations”, which are significant at the 5% and 10% level, respectively. The R^2 is 87%, meaning that 87% of the variability in the target is explained by the regressors, and the adjusted-R^2 is approximately 83%. By analyzing the coefficients in the regression output, it can be observed that each variable has a positive impact on the number of accidents, ceteris paribus, except for the variable “Commute_Automovel_Pass”, which has a negative impact.

[5] Mentioned in the feature engineering section.

In order to compute the residuals[6], this model was applied to each parish data, to obtain the predicted values for the number of accidents, which was then subtracted to the registered values. The residuals were then plotted in a map with the parishes of Lisbon, in Figure 9:

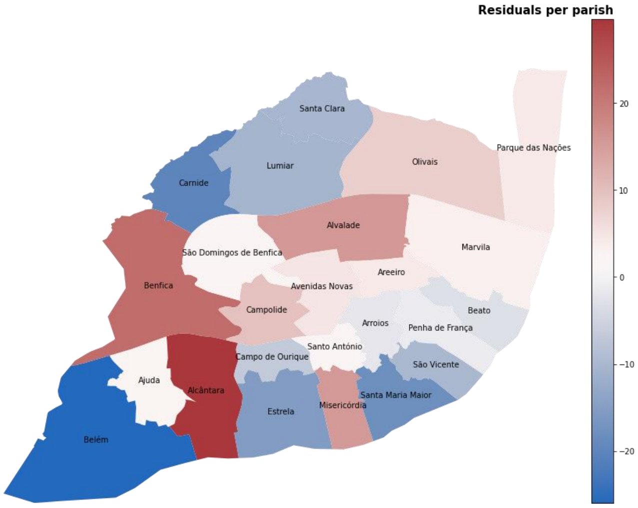


Figure 9 *– Residuals plotted by parish*

By observing the residuals, the majority of the parishes have small deviations from the model, below 10 accidents (absolute value). When observing Figure 9, it can be noticed that Belém stands out as one of the parishes where the model predicts more accidents than actually happened. On the other hand, Alcântara and Benfica stand out for having negative deviations, meaning that the model predicts less accidents compared to the observed.

b. Discussion

In order to evaluate the quality of the dataset that was provided by the Lisbon Municipality, the number of accidents was compared with the results reported in 2019 about road safety published by ANSR. In this report 8.232 accidents were registered, however, in the dataset that was sent, there are only 2.769 accidents during the same year, therefore, probably this data corresponds to a sample.

Other than that, the percentage of missing values was negligible and the observations of the several variables were uniform and standardized within each field (possibly the information was collected using predefined possible values for the variables).

When analyzing the evolution of the number of accidents during the year of 2019 (Figure 1), it is possible to observe that the number of accidents remained fairly constant during the year.

As per the factors most associated with the occurrence of accidents, apparently most accidents occurred in the presence of favorable external factors. Contrarily to what would be anticipated, the higher number of accidents did not occur in intersections or roundabouts. Additionally, it is also verified that the higher proportion of accidents occurred in dry weather and dry road conditions. This data leads us to believe that accidents may therefore be more influenced by human error rather than by external factors.

[6] The residuals are calculated by subtracting the real number of accidents and the number of accidents predicted by the model.

As such, the focus should be instead on the parishes. In terms of number of accidents (Figure 4 and Figure 5) it is clearly possible to observe that some parishes are more prone to the occurrence of accidents, namely Avenidas Novas, Alvalade and Lumiar. Also, the most prevalent types of accidents are collisions and runovers, which is consistent with the type of environment (metropolitan area) we are analyzing. Generally, the parishes with higher frequency of accidents also rank higher when evaluating each type of accident separately. However, among the parishes with lower number of accidents, Santa Maria Maior and Estrela stand out by the number of high runover accidents and overturn accidents, respectively. Santa Maria Maior is a parish with several pedestrian areas, therefore more prone to the occurrence of runovers, while Estrela's predominant narrow streets and steep inclination could explain more overturn accidents.

Since the number of accidents in each parish may be influenced by the circulation volume and other factors, these characteristics have to be taken into account in order to normalize the number of accidents and assess if their ranking is not being influenced by endogenous characteristics.

As such, a regression model was built to account for the influence of these factors and be capable of explaining the volume of accidents in each parish.

In order to build a robust model, variables that are able to mimic the traffic in the parish (such as the number of people commuting), the complexity of the parish (such as the number of intersections and metro stations) and the size of the parish (such as the dimension, the population and the housing prices[7]) were considered.

All the variables included in the model positively impact the occurrence of accidents in the parish, except one: the number of passengers[8] commuting by car. This may be explained by the fact that sometimes the passengers may help the driver and therefore assist in preventing the occurrence of accidents.

Nonetheless, the model may have limitations that can influence the results. Since the results were aggregated by parish, only 24 observations (the number of parishes under analysis) were used to build the model. Also, the data collected was from the most recent census available (2011), while the accidents in the dataset were registered during 2019. For simplicity, it is assumed a constant demographic renovation, allowing the comparison of the two data sources (assumed that the distribution of the variables between the different parishes has remained approximately the same). In the future, following up on this research, it might be interesting to expand the observations to include more years of data and account for time and seasonal effects.

For the majority of parishes, the number of accidents estimated by the model is very similar to the ones registered in the dataset. This allows to conclude that for those parishes the endogenous characteristics greatly influence the number of accidents registered. Therefore, parishes like Avenidas Novas and Alvalade may have a greater number of accidents simply because they are densely populated or have more traffic. Nevertheless, for Belém and Alcântara, the model is over and underestimating the number of accidents, respectively. This means that there may be other factors other than the considered controls, that are not being taken into account in the regression, which impact the occurrence of accidents.

[7] The house pricing variable was used to try and explain the dimension of the parish, since lower house pricing is usually associated with a higher population density.

[8] As opposed to drivers, which were removed due to the high correlation with other features.

IV. Conclusions

Accidents, severely impact the urban livability. As such, this project aims at answering to a challenge posed by the Lisbon Municipality to identify the areas with the highest incidence of accidents and their correlation with external factors. From an early exploratory analysis, most accidents occur in the presence of favorable external factors, which leads to the belief that they may be more influenced by human error. Nevertheless, a regression model was built to account for the influence of endogenous factors and explain the volume of accidents in each parish. This allowed to compare the real number of accidents registered with the predicted values for each parish and identify the ones with a misadjusted number of accidents for their endogenous characteristics. For most parishes, the residuals are negligible, therefore allowing to infer that accidents occurring in Avenidas Novas and Alvalade are related with the endogenous factors of the parishes and therefore it cannot be implied that these parishes are more dangerous than the others. However, other parishes like Belém and Alcântara have high residuals, indicating that there may be some characteristics explaining the occurrence of accidents other than the ones contemplated in the model. As such, these additional factors should be monitored closely, so they may be applied to other parishes and reduce the number of accidents.

V. Acknowledgements

The authors acknowledge the support for this research of the LxDataLab (Lisbon Urban Data Laboratory) for kindly providing the necessary datasets and for the partnership between the Lisbon Municipality and Nova -Information Management School. Also, a special thanks to professors Flávio Pinheiro and Vítor Manita, for the mentoring and encouragement. An additional thanks to professor Flávio Pinheiro for the brainstorming session that allow us the grab a new course for the project.

VI. References

(1) https://lisboainteligente.cm-lisboa.pt/lxdatalab/desafios/identificacao-de-pontos-de-incidencia-dos-acidentes-rodoviarios-e-da-sua-correlacao-com-outros-fatores/

(2) https://geodados-cml.hub.arcgis.com/datasets/freguesias-2012?geometry=-9.423%2C38.692%2C-8.901%2C38.785&selectedAttribute=Shape__Area