Lisbon accident evaluation: Which factors lead to accidents and wounded drivers and passengers

Fábio Oliveira[1], Gonçalo Duarte[2] & Válter Frade[3]

[1]Fábio Oliveira, m20201057, m20201057.novaims.unl.pt
[2]Gonçalo Duarte, m20201329, m20201329.novaims.unl.pt
[3]Válter Frade, m20201052, m20201057.novaims.unl.pt

Abstract: Everyday lives are lost due to car accidents. However, by studying the influence of driver, road, traffic, and atmospheric characteristics we hope to better understand what influences the occurrence of accidents raising awareness and reducing their occurrences. For this study we combined statistical analysis – lasso method and spearmen correlation – with Uber's H3 grid classification system for geographic representation and to better understand what influences the occurrence of accidents. Through these methods we were able to identify factors that correlate positively to the occurrence of accidents - number of zone crossroads and traffic lights – as well as highlight 3 main zones with a relatively high accident incidence. We were also able to identify several factors that lead to wounded drivers and passengers, namely the type of vehicle involved and age of the driver and passenger.

I. Introduction

Motor vehicle accidents can have a high impact on our everyday life, not only regarding public infrastructure but also when we consider the persons involved. Solely in Portugal, there has been an average of more than 30 thousand accidents per year that cause more than 40 thousand wounded and 400 deaths [1]. For this reason, the study of the main factors leading to vehicle accidents is of the most importance for many cities and countries around the world.

However, accident occurrence is a complex problem, with many factors in play. From driver characteristics, such as skill level, experience, risk-taking behaviors, and age, to external conditions such as quality of the road, traffic, atmospheric conditions number of persons in the vehicle, and vehicle type.

Several studies have been made where different factors are evaluated [2, 3], however, this type of evaluations can present some limitations and difficulties. For example, in Portugal, you have the option of reaching a mutual agreement when an accident occurs. This type of accident never reaches the police and fireman reports and can lead to a non-representative sample of the accidents that occurred in the city if only the reports are evaluated.

Another limitation is the gathering of information regarding the accidents. Although most of the external factors can be obtained and evaluated, some of the driver's characteristics are difficult to obtain. How can one evaluate if a driver is inexperienced for example? Or if the driver was distracted before the accident?

Nonetheless, and even with these limitations one can evaluate the police and fireman accident reports and try to reach some conclusions, being that the objective of this work.

II. Data

a. Description and Extraction

For this study, we used several datasets supplied by LxDataLab:

- *Accidents in Lisbon registered in 2019 by the "Autoridade Nacional de Segurança Rodoviária" (ANSR)*: Dataset with multiple information regarding the accident characteristics, vehicles and persons involved.

- *Accidents in Lisbon registered in 2019 by the "Regime de Sapadores Bombeiros" (RSB)*: Dataset with information regarding the accident location, type and date.
- *Lisbon street height, slopes, crossings, and traffic lights shapefiles.*
- *Lisbon traffic jams in 2019*: Information regarding the city traffic registered by Waze in the year 2019.

b. Transformation

We started by evaluating the accident data consistency for errors and misclassifications[1]. The main objective was to create a single accident dataset. However, both files had different information and geolocation missing values, so in the end we divided the accidents information into three datasets: accidents without geolocation, accidents with geolocation and accidents with ANSR report information.

Regarding traffic information, several transformations had to be performed. Namely, the conversion of milliseconds to date, and the calculation of average delays per hour, month and street.

Also, both shapefile's data and traffic information were aggregated based on latitude and longitude values using the python package "h3" developed by Uber [4] which divides space into hexagons and allows the creation of a spatial grid. For this division we selected a hex size of 8 to have the best compromise between hex size and information per hex.

Lastly, Geopandas [5] and folium [6] were the libraries used for reading and visualizing geographic information.

III. Results and Discussion

a. Accident occurrence – geographic analysis

We started our evaluation by counting and visualizing the number of accidents that occurred in Lisbon by created *hex* cell.[2]
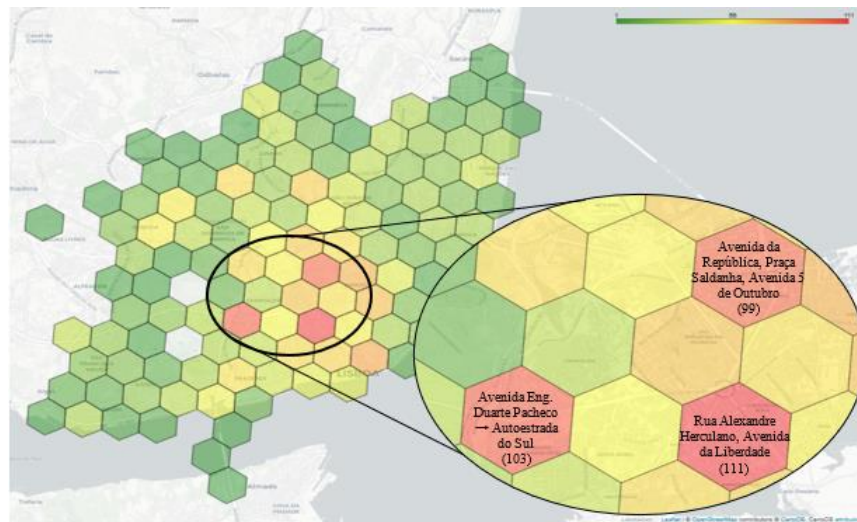

Fig. 4: Geographical representation of the h3 hex cells, with resolution=8, colored by the number of accident occurrences that ranges from 1 (green) to 111 (red). Zoom in on the *hexs* with the most accidents.

---

[1] Check Jupyter notebook: 0.0-data-preparation.ipynb

[2] See data transformation

The three highlighted red *hexes* represent the zones with the highest number of accidents in Lisbon and the respective most known streets in each hex.

A classification of the highest accident incidence zones, depending on the type of accident, was also done but not included in the final report.[3]

b. Crossroads, traffic lights, elevation, and slope analysis

Following the aggregation of the accident data we evaluated road characteristics impact on vehicle accidents. For the calculation of an indicator of crossroads/ traffic lights per *hex* we decided to count the number of crossroads in each defined *hex* and then analyze it versus the number of accidents. The following plots were obtained:
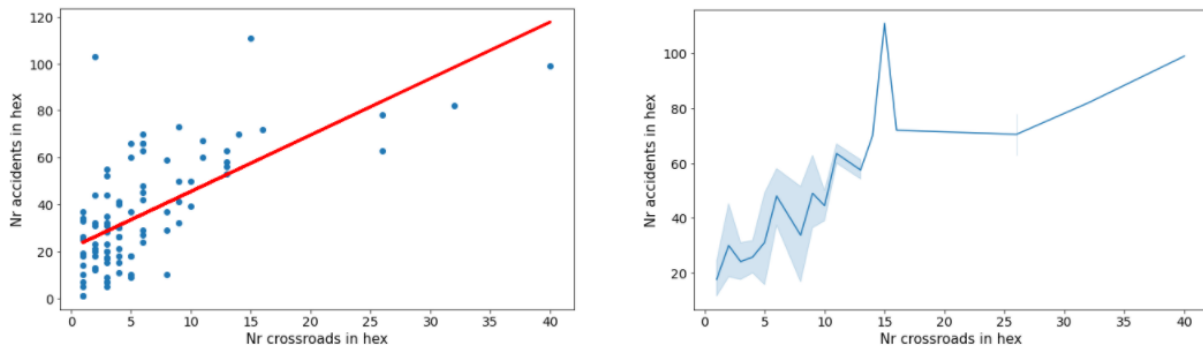

Fig. 5: Scatter and line plot of the number of accidents for each hex and the respective traffic indicator with the best linear fit.

The previous plots - associated with the spearman test results of 0.63 with 1.67×10^{-11} p-value (inferior to 0.05) - show that there is a moderate correlation between the number of crossroads and accidents in a zone. This positive correlation is somewhat expected due to the nature of crossroads which may lead to accidents when traffic laws are not respected. The altimetry shapefile provided elevation values for different points of Lisbon which were averaged for each hex and compared to number of accidents. This resulted in an apparently random scatter plot and low correlation value. A similar procedure was done for the slope, but similar results were obtained.[3][4]

c. Traffic Characterization

Using data provided by Waze, we also evaluated the traffic in Lisbon and its influence on vehicle accidents. We started by grouping the average traffic delay per hour and month so that we could obtain a global traffic profile for Lisbon.

---

[3] Check Jupyter notebook: 0.3-accident-geographic-view.ipynb
[4] Check Jupyter notebook: 0.4-traffic-data-processing.ipynb
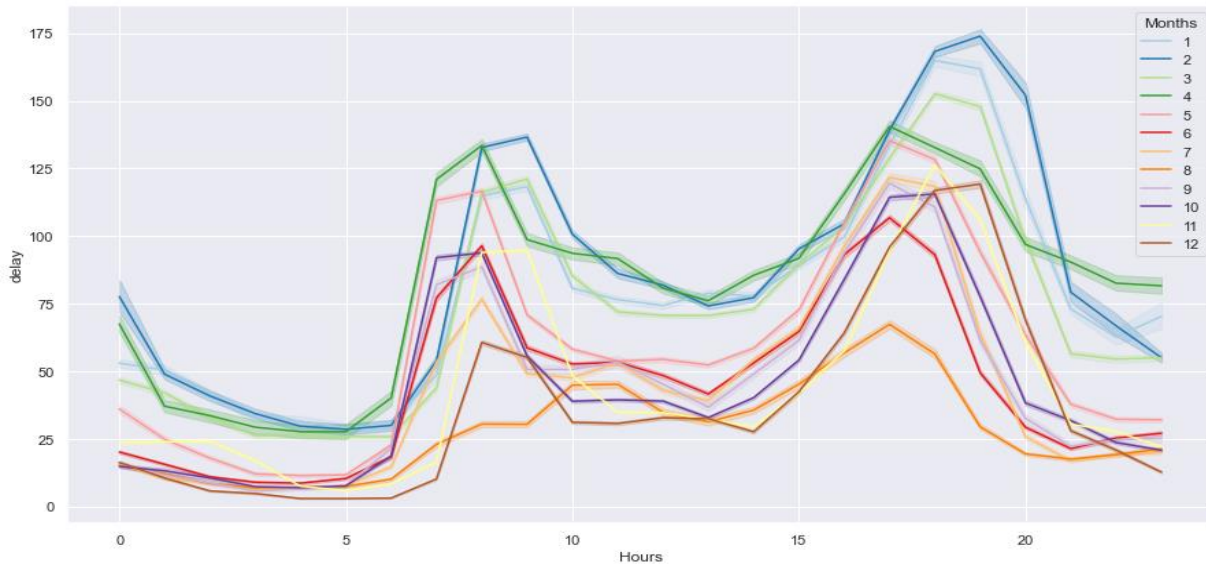
Fig 7. Hourly profile of Lisbon traffic grouped by month.

On one hand, we can see that we have two major traffic peaks around 8 and 18 o'clock, corresponding to Lisbon rush hours. On the other hand, the hours with the lower traffic values are between 4 and 5 o'clock. We can also see that August is the month with lower levels of traffic while January and April are the ones with higher values. [5]

A similar procedure to the road characteristics was also performed for the traffic. We associated to every accident a traffic value corresponding to the previous 10 minutes and then evaluated the correlation between the average traffic and accidents in a zone. This also resulted in an apparently random scatter plot and low correlation value. [6] [7]

d. Accident characteristics leading to wounded drivers and passengers

Besides the accident evaluation and their main causes, we also tried to evaluate what were the main accident characteristics that lead to wounded drivers and passengers. [8]

We approached this evaluation as a classification problem were all the wounded passengers (lightly, badly wounded and death) where considered a positive outcome and using the accident characteristics as predictive features. We decided to apply Lasso regression in this classification model due to its simplicity and because this regularization method tends to convert the coefficient of the variables that are less significant to 0. In the end we obtained the coefficients presented at the next graph (note that all the coefficients were significant).

---

[5] Check Jupyter notebook: 0.4-traffic-data-processing.ipynb
[6] Check Jupyter notebook: 0.3-accident-geographic-view.ipynb
[7] Check Jupyter notebook: 0.4-traffic-data-processing.ipynb
[8] Check Jupyter notebook: 0.2-accident-police-evaluation.ipynb
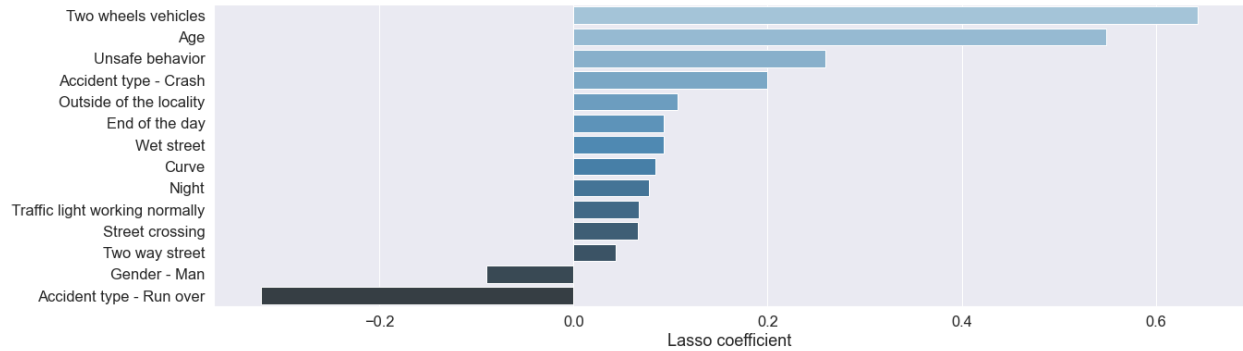
Fig. 3: Lasso coefficients for different factors

As can be seen most of the factors seem to have a positive impact on the occurrence of wounded.

- The two factors with a greater positive impact are if you travel in a two-wheel vehicle or if you are older. Intuitively, this makes sense since when you travel in a two-wheel vehicle you tend to fall and need medical assistance. At the same time if you are older, you can have more health complications.

- The remaining factors that impact positively the occurrence of wounded are if an accident is a crash, if it occurs at the end of the day or during the night, if the road is wet, if it occurs outside of the localities, in a curve in a two-way street or in a crossing.

- Regarding, the factors that negatively affect the occurrence of we have if the driver or passenger is a male of if the accident type is a run-over, which makes sense, since if you run over a person as a driver or passenger you are less likely to be wounded.

Although, many conclusions can be taken from this evaluation further studies need to be performed to estimate the real impact of these factors.

IV. Conclusions

Although intuitively we know that many factors influence the occurrence of accidents in our study, we could only prove that the occurrence of accidents depends on the number of crossings and traffic light. However, we can't say that the other parameters don't influence the occurrence of accidents, we just couldn't find any relation in this study. Regarding driver and passenger wounding, we were able to prove that some accident factors increase this probability such as the type of vehicle and road characteristics while others reduce it such as the type of accident and gender of the driver/passenger..

In conclusion, we hope that this work will serve as a first approach for further studies.

V. Acknowledgements

VI. References

[1] Pordata:
https://www.pordata.pt/Portugal/Acidentes+de+via%C3%A7%C3%A3o+com+v%C3%ADtimas++feridos+e+mortos+++Continente-326

[2] AshrafI, Hur S, ShafiqM, Park Y (2019) Catastrophic factors involved in road accidents: Underlying causes and descriptive analysis. PLoSONE 14(10)

[3] Rolisona J., Regeva S., Moutarib S., Feeney A. (2018) What are the factors that contribute to road accidents? An assessment of law enforcement views, ordinary drivers' opinions, and road accident records. Accident Analysis and Prevention 115 (11-28)

[4] H3 library: https://pypi.org/project/h3/

[5] Geopandas: geopandas · PyPI

[6] folium: folium · PyPI

[7] uber github: https://github.com/uber/h3-py-notebooks/blob/master/notebooks/urban_analytics.ipynb