Accident analysis in Medellín

Juan Diego Pineda-Jaramillo[1,+], Javier Moreno[2,+], and Esteban Silva-Villa[3,+]

[1]Ferrocarril de Antioquia, Technical department, Medellín, COlombia
[2]Owens-Illinois, IT department, Medellín, Colombia
[3]Univesity of Antioquia, Physics Institute, Medellín, Colombia
[+]these authors contributed equally to this work

ABSTRACT

One of the most complicated problems in large cities is traffic. There are many possible causes related with mobility accidents, such us weather, light, human responses, among others. Trough the analysis of a dataset of accidents happened in Medellín, Colombia, we studied the time variation and spatial correlation of all accidents between the years 2014 and 2016. Our results show that the main highway that crosses the city has a large probability of accidentallity. Using a larger same, we developed a method to estimate the probability of the road infrastructure to influence the accidents, and how to implement those solutions to decrease the probability of mortallity. Our recommendation is to reduce the speed limit with road controls and to increment light to improve visibility.

1 Introduction

One of the world's leading causes of death and injuries is urban traffic accidents. According to the World Health Organization (WHO), traffic accidents was the leading cause of death for children and young adults aged 5-29 years, and the eighth leading cause of death for all age groups, surpassing HIV/AIDS, tuberculosis and diarrheal diseases in 2018 (WHO, 2018). In addition, more than a million people die each year on the world's roads, and 90% of those deaths occur in low- and middle-income countries that represent 82% of the world's population (WHO, 2018), (Wegman, 2017). For this reason, the United Nations declared 2011-2020 as a decade of action for road safety (United Nations, 2010), with the aim of reducing the number of deaths of traffic accidents around the world through national and citywide plans, considering five key strategies (United Nations, 2011):

- Road safety management
- Safer roads and mobility
- Safer vehicles
- Safer road users
- Post-crash response

These key strategies are focused on security and prevention, related to strategies in the medium and long term, where the commitment of nations is a key success factor to achieve the goal of the decade, through the design of management plans, including investment, transfer and creation of knowledge to face the current problem, and to implement sustainable actions in the future.

Colombia is a middle-income country with 13.8 deaths in traffic accidents per 100.000 inhabitants in 2018, where the rate for high-income countries is around 9. On top of this, traffic accidents were the second largest cause of violent death in Colombia for the same year, with 26.7% of the total number of deaths (Instituto Nacional de Medicina Legal y Ciencias Forenses, 2011).

Particularly in Medellín, which is the second largest cities in Colombia, with more than 2.5 million inhabitants (and almost 3.8 million inhabitants in its Metropolitan Area), there has been an increasing trend in the number of traffic accidents. According to 2014 statistics by the Municipality of Medellín (Alcaldía de Medellín, 2014), from 2008 to 2014, the number of accidents grew up to 20.14%. One of the main causes of this problem is the growing number of vehicles, which from 2008 to 2014 increased from 767.548 to 1.234.946 (i.e., by 60.9%).

This project is looking for optimal strategies with the aim of reducing the traffic accidents in Medellín. It is necessary to analyze different characteristics in the traffic accidents in the city to this goal. Hence, in this project we want to analyze

the trends in the traffic accidents in Medellín, and uncover which are the most influential variables that cause accidents (i.e. pavement state, weather conditions, driver's age, location, hours, days, among others), and which variables have more trend to conduct to fatal accidents.

The main contribution of this project is to perform a detailed analysis of the accidents' factors in Medellín, and build a tool for the Public Administration to conduct strategies to reduce these traffic accidents.

2 The data sample EDA's

2.1 The sample data set used

To achieve our goal, we have the a data sets collected by the municipality government. The original data set contained personal information regarding the people involved in the accidents registered. We modified the sample by making the information totally anonymous.

1. *Inf_General_y_Caracteristicas_Vias.xlsx*
   This data set contains details of general characteristics related to all the accidents occurred in Medellín from January 2009 to November 2016. Bellow we present a brief summary of the attributes contained in this data set:

   - Accident ID
   - Address and location of the accident (with no coordinates)
   - Date and time of the accident
   - Type of accident, including if the vehicle crashed with a fixed object or another vehicle
   - If the accident was in an intersection, in a street or other type of infrastructure (i.e. viaduct, tunnel)
   - Weather conditions
   - Geometric parameters of the street, as one or two-way street, signaling, number of lanes, pavement state, light traffics, visibility, illumination, among others

2. *AccidentalidadVictimas.xlsx*
   This data set contains details of the victims (injuries or deaths) related to all the accidents occurred in Medellín from January 2009 to December 2016. Bellow we present a brief summary of the attributes contained in this data set:

   - Accident ID
   - Address and location of the accident (with no coordinates)
   - Date of the accident
   - Different characteristics of the traffic accident victim, such as nationality, birthday, gender, victim type (passenger/ driver, pedestrian), type of accident (injury, death)

3. *Conductores_vehiculos.xlsx*
   This data set contains details of the vehicle drivers related to all the accidents occurred in Medellín from January 2009 to December 2016. Bellow we present a brief summary of the attributes contained in this data set:

   - Accident ID
   - Address and location of the accident (with no coordinates)
   - Date of the accident
   - Different characteristics of the traffic accident driver, such as nationality, birthday, gender, type of accident (injury, death, just damages), driver' license, public transport driver (Y/N), type of vehicle, failure of the vehicle(brake system, direction system, tyres, lights, vehicle whistle, vehicle suspension, None)
   - Hospital where the victims were taken (if injuries or deaths)

- Description of injuries

4. *Accidentalidad_georreferenciada_*2014 – 2019.*csv*
   This data set contains details of the generalities related to the location of all the accidents occurred in Medellín from January 2014 to June 2019. Bellow we present a brief summary of the attributes contained in the data set:

   - Accident ID
   - Address and location of the accident (with coordinates)
   - Date and time of the accident
   - Type of accident
   - If the accident was in an intersection, in a street or other type of infrastructure (i.e. viaduct, tunnel)

   In this project we joined all the data sets mentioned above, and after (geo)located each point over a map, we analyzed the trends in the traffic accidents in Medellín, studied which are the most influential variables that cause these and which variables have more trend to conduct to fatal accidents.

2.2 The exploratory data analysis (EDA) on the sample

For each data set mentioned above we removed rows that do not have complete information that can be useful for this study. Once the files have the relevant information, we matched the files using the accidents ID recorded by the agents at the location where each situation happened.

The original files have a date range covering from 2009 to 2016. However, the most complete sample we relied on covers the years from 2014 to 2016. This last time range was the one used for the study to make the visualization of the actual situation in the city. This is because it is the most reliable data set, but also, because is the most recent one, allowing to understand better the actual situation.

The final data set consist of a sample of over 100.000 data points with a full set of information.

3 Methods

3.1 Data presentation

The results of the data study on Medellín showed interesting, although not surprising, indications about traffic in a large city.

As can be observed in figures 1, 2, 8 and 3, there is a large number of accidents happening in a short time interval.

It is important to highlight the results for the day Friday. Not only by hour, but also by month, this particular days presents the worse results, showing a large number of accidents happening consistently in a 2 year time range. These results are added up for all possible causes of accidents in the data set, but can be observed as separate results, analysing each possible severity type independently (results observed but not shown here).

Regardless of the results for the day Friday, there is a consistent results showing that, during the week days, driving between 11am and 7pm, the possibility of an accident is high.

3.2 Finding the danger hotspots for a given route

Among many possible applications, we decided to develop a solution that, given a route, would identify what are the most dangerous hotspots and qualify the danger level based on the amount of accidents that have occurred during the last year of data (i.e. year 2016). To achieve this, we started by filtering the dataset containing the coordinates of the accidents to include only the events from July 1st 2018 to June 30th 2019, this gave us a total of 39715 events.

Using scikit-learn DBSCAN (Density-based spatial clustering of applications with noise), by passing the latitude and longitude of the accidents as input parameters and using the harversine metric (which would return the results in units of radians), and considering 6371.0088 kilometers per radian[1], we created clusters of accidents in order to find the hotspots, where at least 20 events were registered in a range of 80 meters. Clusters with less than 20 events are considered as noise.

The resultant number of clusters set contains 477 points, with the center most point of the original dataset as the cluster representative coordinates, and with the count of accidents normalized between 0 and 1 as the threat level (i.e. the threat level is a relative number). Using Google Routes API, by passing the initial and final location of the trip, considering the transportation mode as driving, walking and cycling, and the time of the trip being the moment of the request, the response is parsed to obtain

[1] *https://github.com/gboeing/urban-data-science*, Geoff Boeing

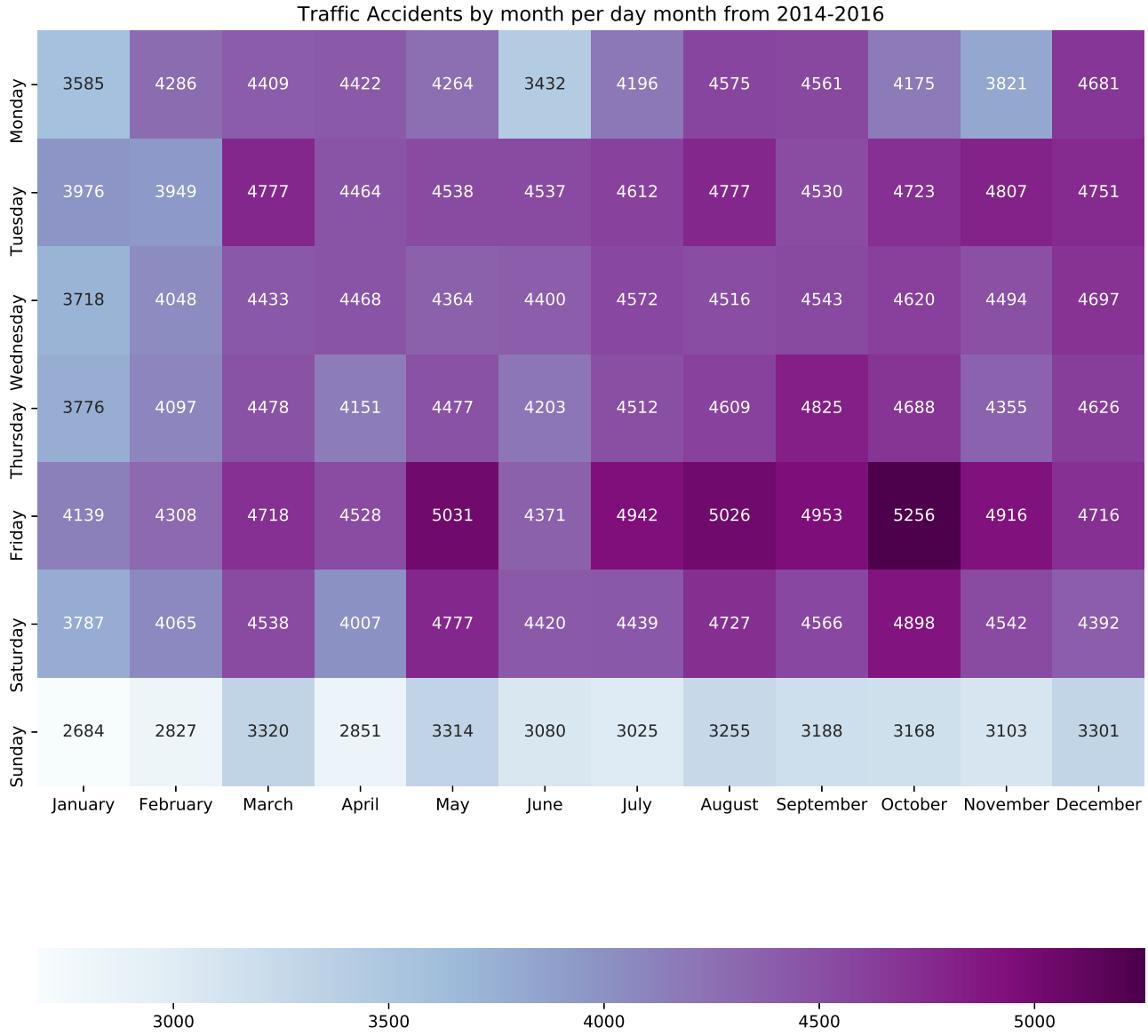

Figure 1. Accidentally by month per day of the week.

the coordinates points of the steps to follow during the trip. Then, all of the points are evaluated against the clusters to find if they are close to a threat cluster (i.e. closer than 80 meters) in order to obtain the threat level[2].

Finally the result is presented as the recommended route by Google plus the hotspots with colors, depending of the threat level. It is important to note that in this scope we did not evaluate points between steps, meaning that if google does not give a hint to the driver during a long segment of the drive, the system is unable to tell if there is any hotspot for accidents in between steps.

3.3 Predicting the severity of an accident based on infrastructure's characteristics

For the public administration, it is mandatory to reduce the number of (serious) accidents in the city. To achieve this, it is possible to implement different strategies that range from creating a road culture in citizens, to perform large infrastructure projects.

To know the possibility of reducing the severity of an accident, through small improvements in infrastructure (mainly), such as the implementation of speed bumps or improving vertical and horizontal signaling on the roads, we used the scikit-learn

[2] *https://github.com/meraldoantonio/AccidentPredictor*,Meraldo Antonio

Figure 2. Accidentally by day of the week per hour.

library. We implemented different Machine Learning algorithms to classify almost 360 thousand accidents over 8 years into serious or non-serious based on different infrastructure parameters of the location where the accident occurred.

To select the variables that can be passed to the model, we observed the possible correlations (or their lack there off) in order to reduce the number of free parameters. THe correlations observed are presented in figure 5

After testing different Machine Learning algorithms such as Logistic Regression, Linear Discriminant Analysis, Decision Trees Classifiers, Gaussian Naive Bayes and Random Forest Classifiers, we obtained the following results applying cross-validation after splitting training data in 10 kfolds:

- Logistic Regression: 60.29%
- Linear Discriminant Analysis: 60.73%
- Decision Tree Classifier: 79.37%
- Gaussian Naive Bayes: 79.33%
- Random Forest Classifier: 79.33%

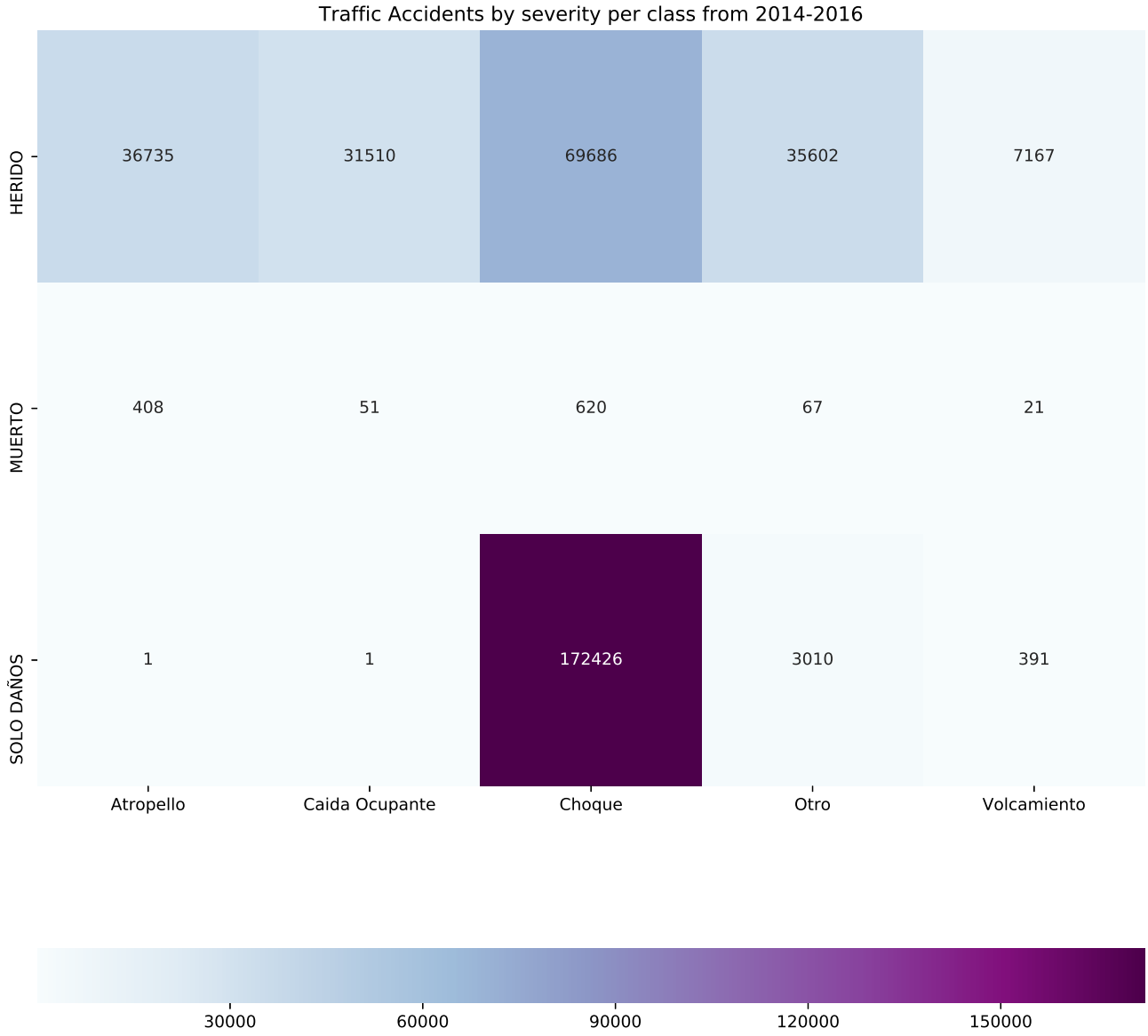

Figure 3. Accidentally by class per severity type.

Finally, we evaluated the decision tree and the random forest, achieving better results on accuracy for testing data with the Random Forest.

So, for the chosen model, the Random Forest Classifier, we got an accuracy of 79.16% in training data, and an accuracy of 79.37% in testing data. We also used the AUC metric for assessing the model, obtaining an AUC of 0.75. For this reason, we selected this model for deploying in our app.

4 The App

4.1 Backend implementarion

The implementation of all our study was done in *python*.

Figures 7, 8, 9 and 10 are examples of the results obtained with the code created for this project.

Thanks to the implementation of Google Maps API and Google Places, we presented and implemented a complete function in our code that correlate the level of accidentallity (measured through the number of accidents clustered by location) and the possible paths a driver can take to go from point A to point B.

From this point on, Front-end will deploy all predictions and analysis obtained, and will show the potential accident sides

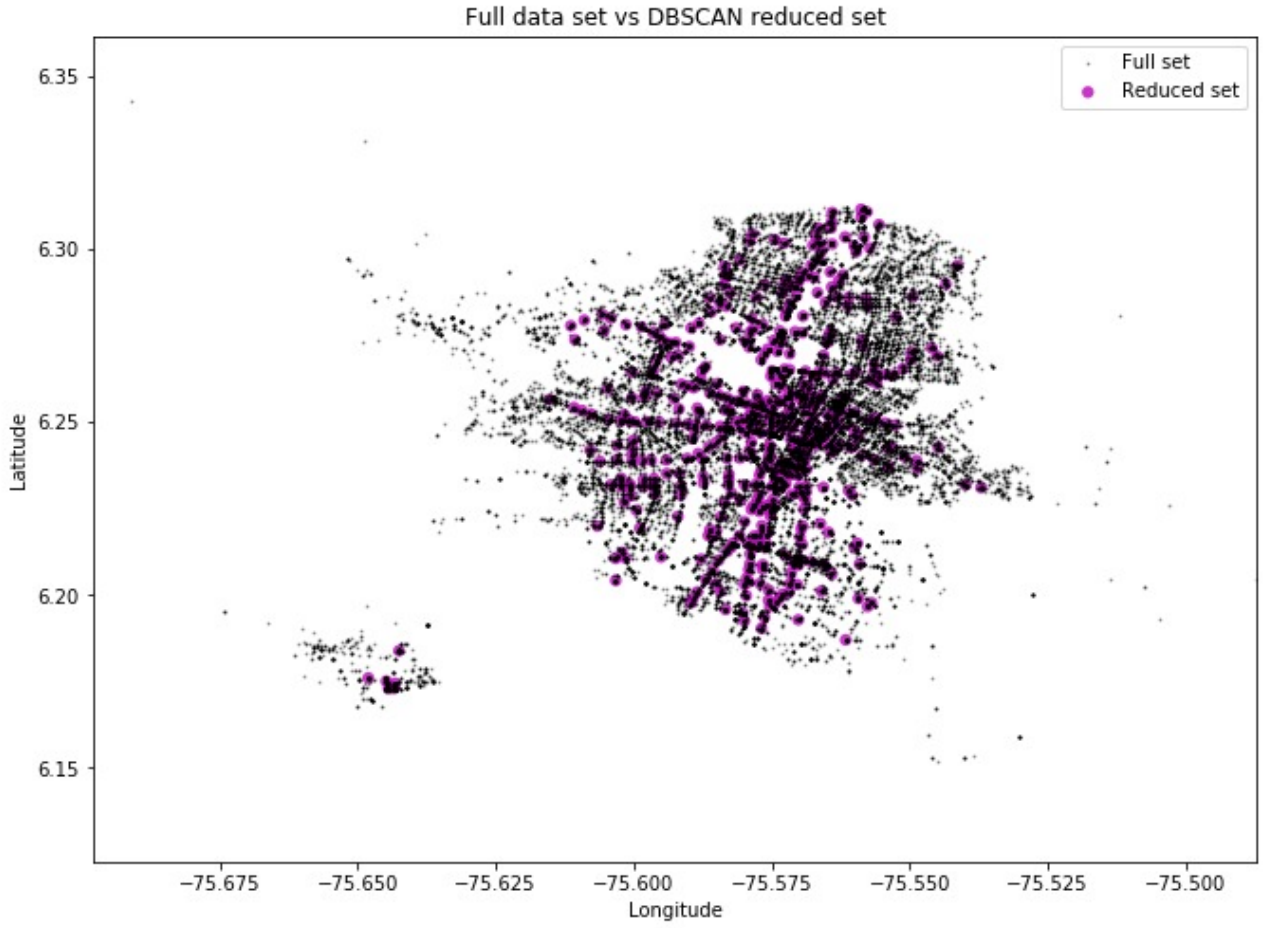

Figure 4. Cluster of accidents generated by the core implemented in our work.

and risks.

4.2 Frontend implementarion

Our project will be implemented in a web page using the AWS services. The website consists of 4 parts.

The first part is a heat-map visualization presenting the data as a Geo-location area over Medellín, where the number of accidents happened during the years studied are shown. This part of the website allows the user to selected among variables of time, days and class of accident.

The second part presents two figures. The first one correlates with the heat-map by showing the variation of the same variables selected. The second part shows the full data set by a correlation among different variables, as shown in the figure (and presented in the website).

In the third part we implemented the Geo-location google API of the clusters of accidents and correlated them with the paths that takes from point A to point B.

The last part shows how the probability of implementing a series of measurements over the streets can decrease the severity of an accident, which in turn decreases the probability of injured and the dead of a person.

5 Discussion and future work

Based on a data set of accidents, and all its especial characteristics, we studied the temporal behavior over the city of Medellín between the years 2014 and 2016.

The results of such study shows that the worse day af the week, regarding accidents, es Friday. While at 5pm is the time of the day that presents the largest amount of accidents, it can be said that between 11am and 7pm, the accidentallity in the city present lasr numbers.

The problem can be approached by spotting those hot spots across the city that present the largest number of accidents. These can be seen in the heat-map for different days of the week (and all of them together if needed), at different times through out the day, separated (or not) by different classes of accidents.

To keep a follow up on these problems, the tool developed on this work could give light on the directions that the administrative government could take in order to reduce the accidentallity in the city.

References

World Health Organization, 2018

F. Wegman, IATSS Research, 40, 2017

United Nations, 2010

United Nations, 2011

Instituto Nacional de Medicina Legal y Ciencias Forenses, 2018

Alcaldía de Medellín, 2014

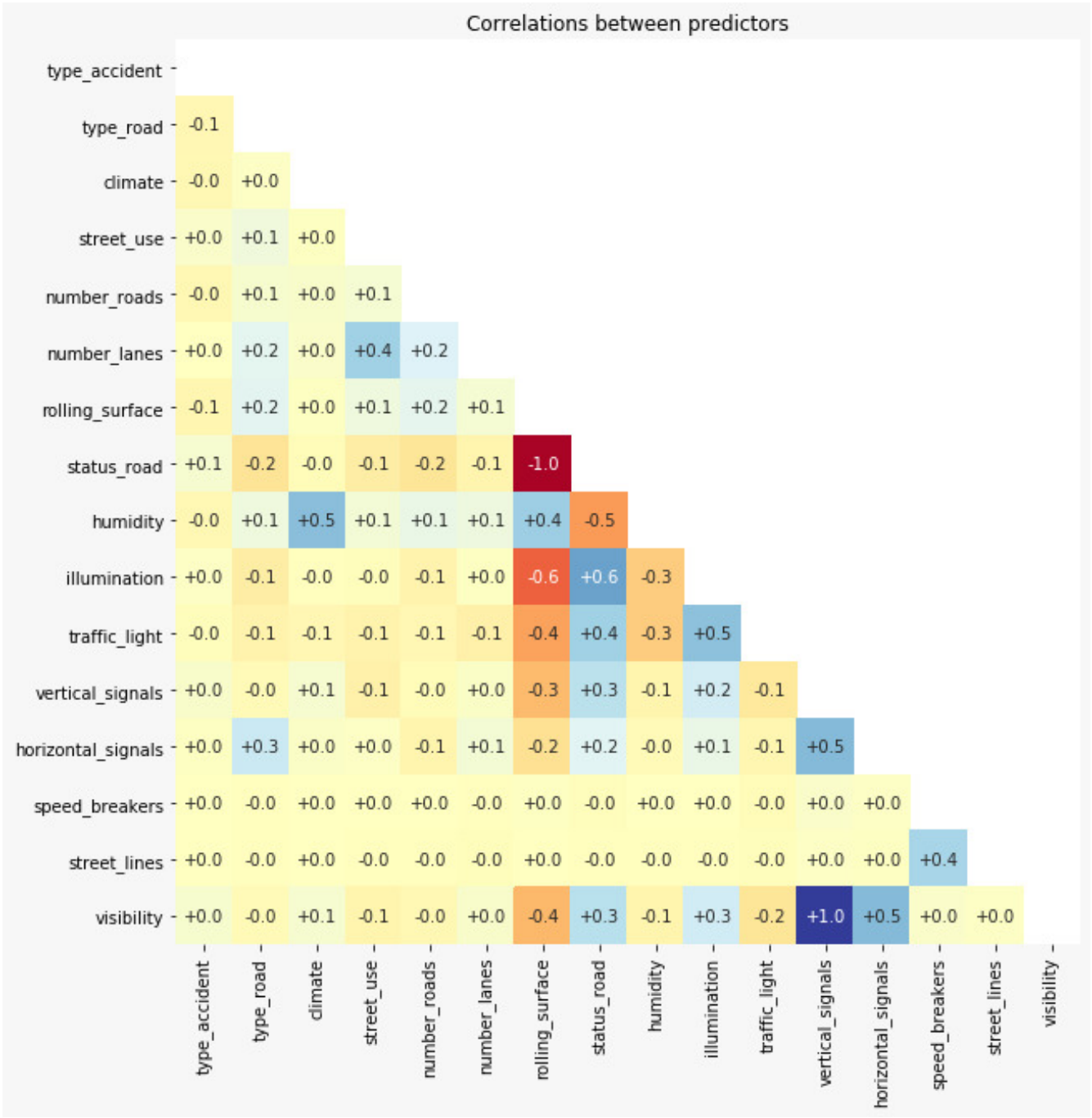

Figure 5. Correlation Between the variables to be used in the model.

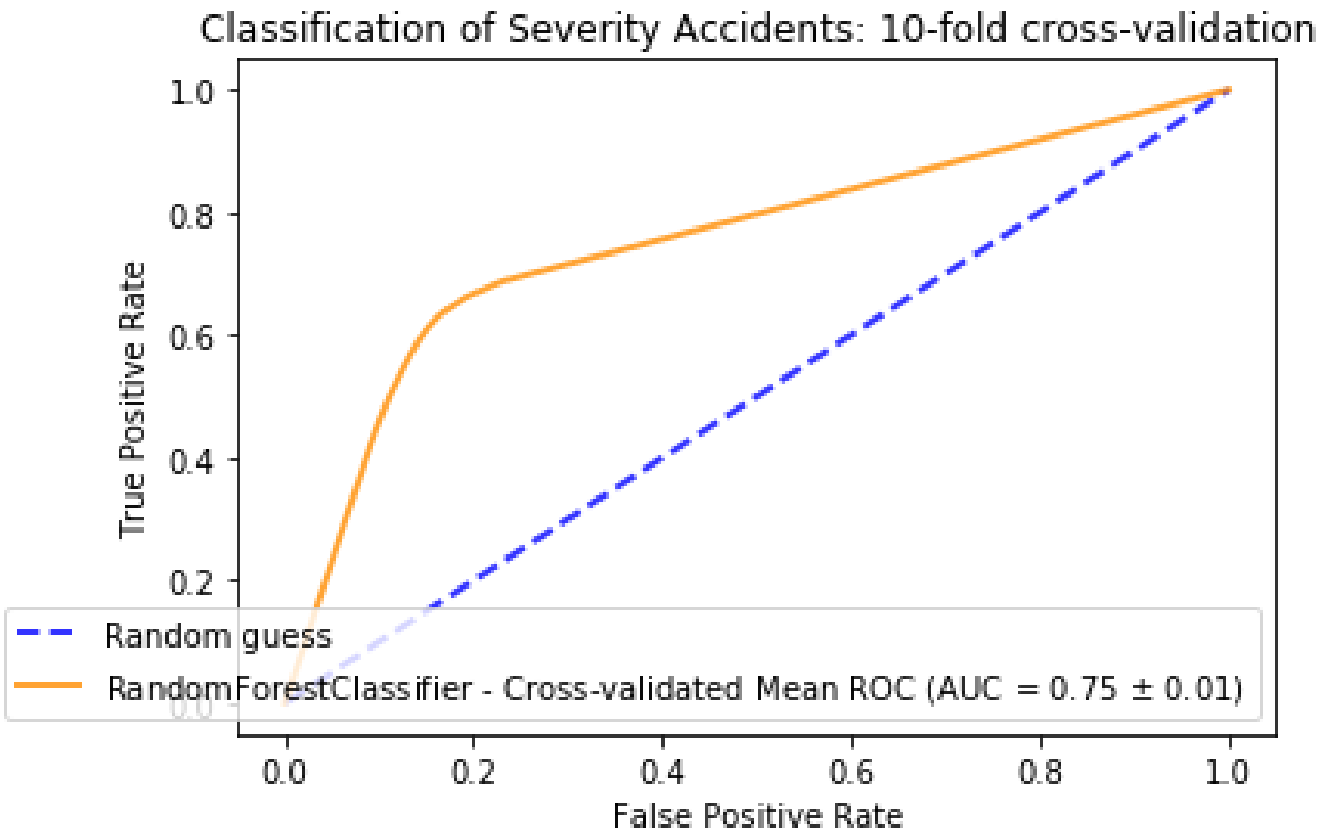

Figure 6. Classification of the severity accidents.

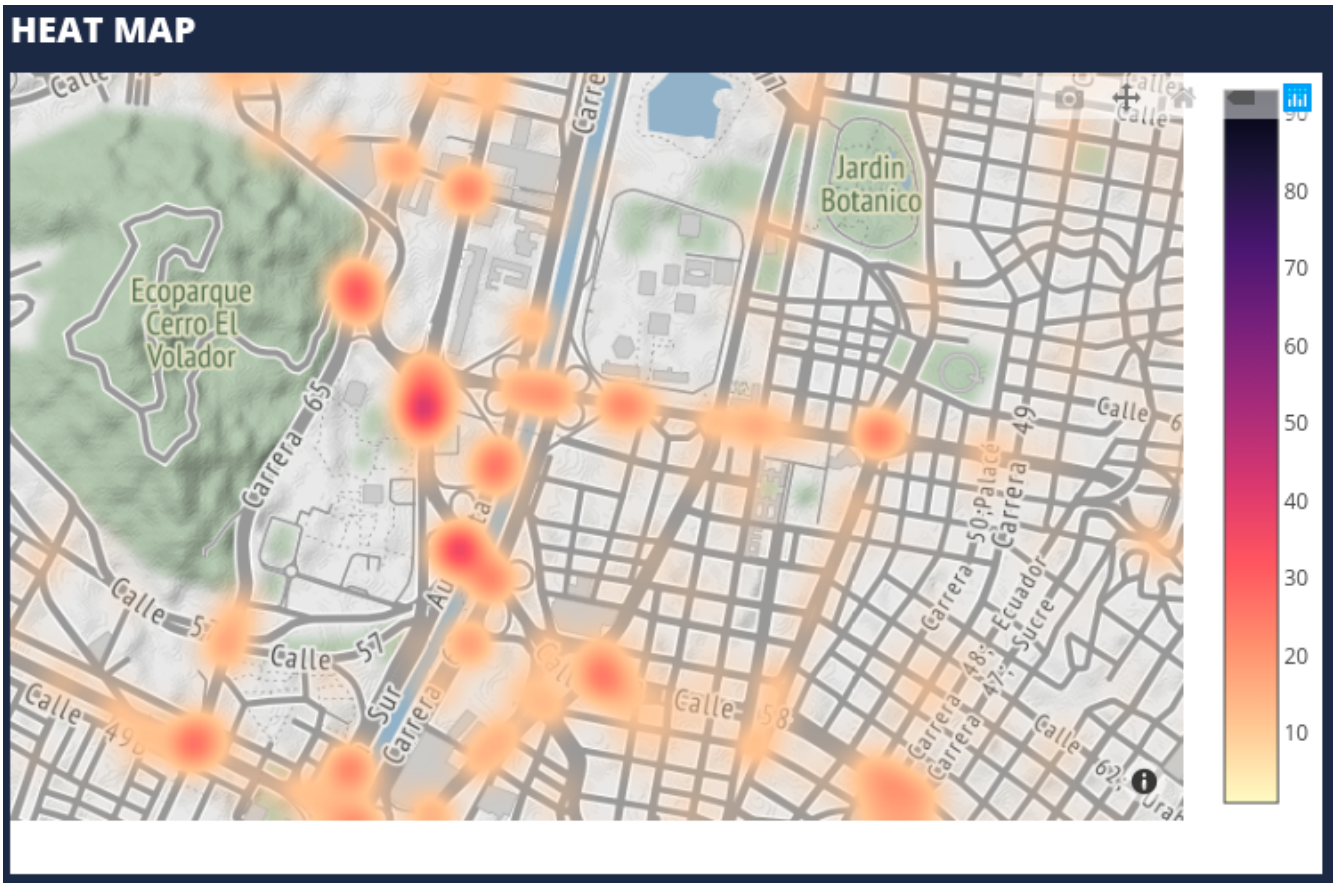

Figure 7. Heat map of the accidentallity in an area of the city.

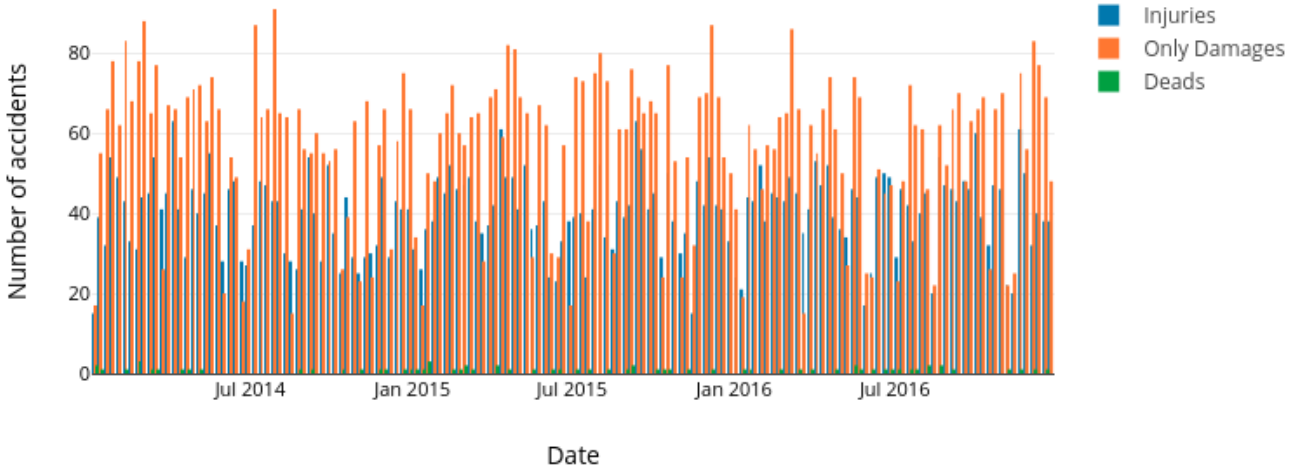

Figure 8. Time variation of the severity of the occurrences of the accidents under analysis.

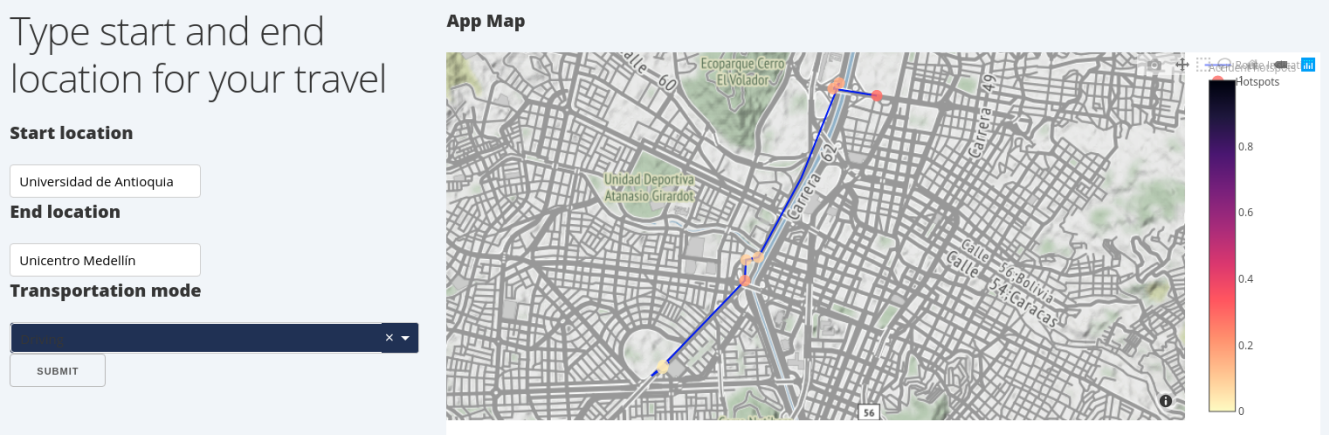

Figure 9. Correlation of route and hot spots for large accidentallity. The selected route shows the path from point *A* to point *B* and the areas of larger probability of suffering an accident.

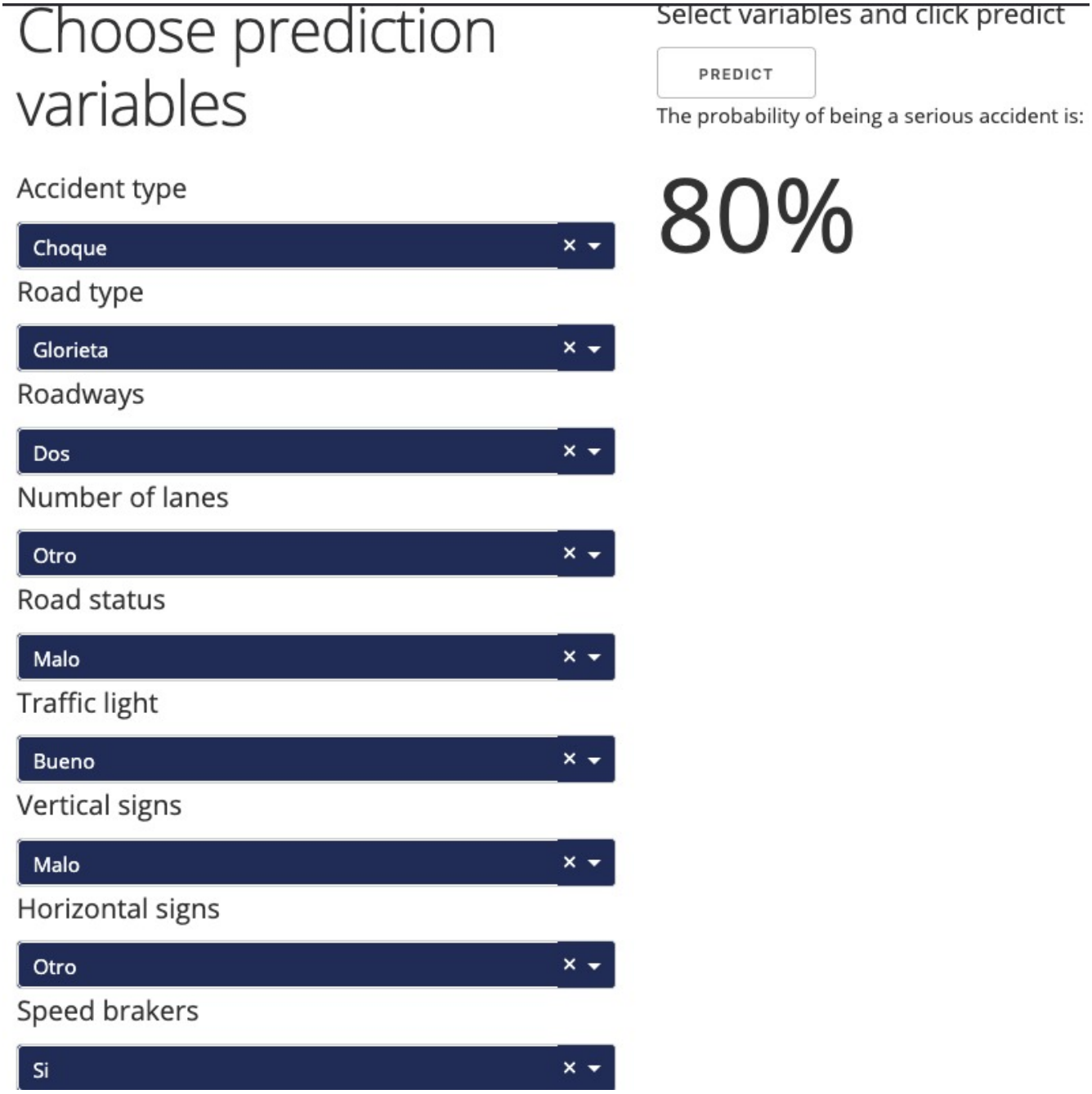

Figure 10. Set of parameters and the corresponding probability based on the proposed model in this work.