Analysis of Incident Variables on Traffic Accidents in Ecuador by using Bokeh library

Ángel Antonio Encalada Dávila

June 18, 2018

Analysis of Incident Variables on Traffic Accidents in Ecuador by using Bokeh library

Ángel A. Encalada-Dávila[1][0000-0001-6259-8464]

[1] Escuela Superior Politécnica del Litoral, ESPOL, PO Box 09-01-5863, Guayaquil, Ecuador
angaenca@espol.edu.ec

Abstract. The Big Data that a lot of public and private institutions have about several variables measured through years, that sometimes it is not harnessed how it should be. The lack of powerful tools for make exhaustive analysis that allows discover insights, patterns, trends, correlations and others important relations has become in a barrier that nowadays could be easily destroyed.
The purpose of this paper is make an analysis about traffic accidents in Ecuador using Bokeh library of Python programming language. This work is routed to discover relations between some variables such as: temperature, precipitations, accident hour and location, victims amount and accident class, that meanly characterize a traffic accident. This study will help to propose new strategies and improve some methods to reduce the accidents amount in Ecuador.

Keywords: Big Data, Traffic Accidents, Insights, Bokeh, Python.

1 Introduction

In the last decades, millions and millions of data have been produced around the world that currently compose the Big Data environment. Likewise, and during this time the analysis of massive data has been enriched, however, this update has not reached all the places that would have been wanted since in the current moments Big Data is still generated but it is not being processing properly, this is, it does not take advantage of or exploit the riches that entails analyzing the massive data. It is necessary to bear in mind that the study of Big Data has achieved over time to be a powerful method to identify trends, insights, patterns and correlations; this contribution has allowed anyone who generates massive data, be it a natural person or a multimillion-dollar company, to have found solutions, strategies, innovations that have solved problems that at first were very complex to solve.

For this study, the central focus is established in the analysis of traffic accident data in Ecuador, which have been collected by the different national institutions responsible for the control of land traffic. The story betrays directly that unfortunately in Ecuador there is no culture to take advantage of the Big Data that occurs on a daily basis. The justifications may come one after the other: a country in development, where the globalization of Big Data has not yet reached society, a country where the lack of data scientists has not allowed this field to leverage, among others; the fact is that the change must start somewhere.

Over time, traffic accidents have been among the leading causes of death, claiming thousands of lives every day, because the chances of survival are often uncertain since they depend on many environmental factors. Several studies have been carried out in recent years in order to find solutions or strategies that help to strengthen measures to reduce the number of accidents and the probability of death, one of them, for example, prediction [1]. This last procedure has been in vogue in recent years, since there are various methods and algorithms to create predictive models: Differential Evolution Algorithm [2], Statistical Modeling, Artificial Neural Networks [3], among others. In the case of Ecuador, innovating in the application of these and other methodologies is still a challenge, since one of the main barriers is not so much the lack of professionals, but the poverty in the reporting of data and measurements, in this case, when traffic accidents occur. The databases have reflected this problem, where the limitations in the registration of information fields are huge. However, this does not prevent studies such as this work from being made, where the information recorded has been taken advantage of and analyzes have been carried out that, besides providing insights, are the basis for further studies.

Bokeh library, a powerful Python tool, has been the fundamental pillar to fulfill the purpose of this paper. Through the application of Bokeh functions and with observation, it has been possible to find some relationships between the analyzed variables that affect traffic accidents.

The organization of the paper follows the sequence: section of introduction to the study of traffic accidents and the use of Bokeh in the data analysis, section of geographic context to understand the worldview in Ecuador around the topic of traffic accidents, section of dataset analysis where pre-research results are reflected, section of methodology that describes the calculation of new study variables for the establishment of correlations, results section and discussion to discuss what has been obtained and execute contrasts over time, section of conclusions and future works on the subject, section of acknowledgments and bibliographical references section.

2 Geographic Context

Ecuador is a country in South America, with an area of around 256,370 km^2. The geographical division of its territory is based on the fault that crosses it from north to south: The Andes Mountain Range. In this way, Ecuador is attributed four well-defined regions: Coast Region, Highland Region, Amazon and Insular Region. The political division is based instead on provinces, that is, territorial jurisdictions with autonomy, but with direct dependence on the central government. Ecuador currently has 24 provinces distributed throughout its territory.

Table 1. Provinces of Ecuador.

Code	Province	Capital	Region
1	Azuay	Cuenca	Highland
2	Bolivar	Guaranda	Highland
3	Cañar	Azogues	Highland
4	Carchi	Tulcán	Highland
5	Cotopaxi	Latacunga	Highland
6	Chimborazo	Riobamba	Highland
7	El Oro	Machala	Coast
8	Esmeraldas	Esmeraldas	Coast
9	Guayas	Guayaquil	Coast
10	Imbabura	Ibarra	Highland
11	Loja	Loja	Highland
12	Los Rios	Babahoyo	Coast
13	Manabí	Portoviejo	Coast
14	Morona Santiago	Macas	Amazon
15	Napo	Tena	Amazon
16	Pastaza	Puyo	Amazon
17	Pichincha	Quito	Highland
18	Tungurahua	Ambato	Highland
19	Zamora Chinchipe	Zamora	Amazon
20	Galápagos	Baquerizo Moreno	Insular
21	Sucumbíos	Nueva Loja	Amazon
22	Orellana	Fco. de Orellana	Amazon
23	Santo Domingo de los Tsáchilas	Santo Domingo	Highland
24	Santa Elena	Santa Elena	Coast

The institutionality in Ecuador is based on having public entities that control the sectors in which a population can be divided. The central point of this study has been the transport sector. A public institution controls the sector from the central government, however, each province has its own transport and roads regulating entity, which does the work done collaborative and systematic.

The database used in the present study is provided directly by the respective transport control entities. It is worth emphasizing that Ecuador has an institution dedicated exclusively to the dissemination of open data, which are of a public nature and are at the service of citizens.

3 Dataset

3.1 General description

The dataset used in the analysis and study of traffic accidents is a file (CSV) with an approximate 30K records. The temporary space in which data collection has been carried out corresponds to two years, 2015 and 2016. The present work has been centered on a detailed examination of the year 2016, while the records of 2015 have served as a means of contrast to assess the changes that have or have not taken over these years.

The registers have several fields of information that have allowed a multiple study of the registered variables. The fields to which you have access are: province, canton, month, day, time, type of accident, cause of accident, area, number of injured, number of deaths and number of victims. The use of the fields as cause of the accident, zone and canton has been exempted due to the absence of other important variables to execute correlations.

3.2 File processing

Before processing the dataset as such, it was necessary to do the lifting of the shapefiles that define the political division of Ecuador in provinces, since the main studies will be carried out by said jurisdictions (see Fig. 1).

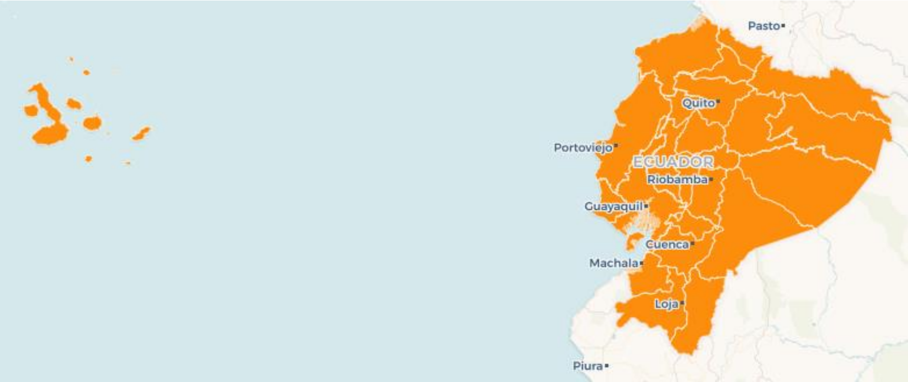

Fig. 1. Political division of Ecuador (provinces).

As a first impression about the information contained in the dataset, a cloud of words was generated with the provinces and the weight of the number of accidents for each one of them; what has been obtained indicates a great disproportion in the weight of the numbers in this field of information (see Fig. 2).

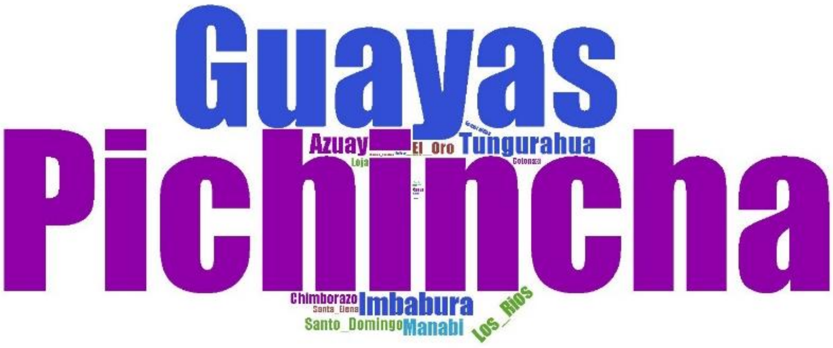

Fig. 2. Accidents per province: "Guayas" and "Pichincha" reflecting a huge disproportion with other provinces.

The most important fields in the dataset are the month, day, time and place of the accident. With these data and with the help of Bokeh, a heat map was created that reflects the number of accidents per province and per month throughout 2016, in such a way that a broader view of the hot spots is obtained where has recorded a greater influx of accidents (see Fig. 3).

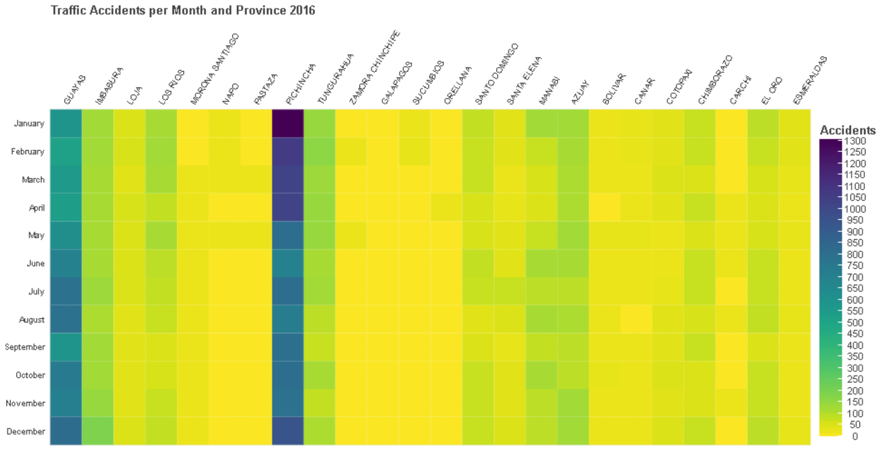

Fig. 3. Heat map of accidents per province and month, 2016.

There is a great relationship between the word cloud and the heat map, where it has been clearly identified that the greatest number of traffic accidents is attributed to Guayas and Pichincha. Accidents range from 600 to +1300; these two provinces lead the ranking being one of the reasons the gigantic population that they have and therefore the great vehicular affluence that occurs day by day.

However, the previous vision is still very broad, that is, it does not allow to discern clearly other approaches that involve traffic accidents. Many jobs, when they have analyzed accident data, consider the days of the week as the main front [4]. A division was made during the week in two important and common regions, the working days (Monday-Friday) and the weekends (Saturday-Sunday) in order to observe some growth, decrease or anomaly behavior, which reflects the choropleths maps (see Fig. 4, Fig. 5).

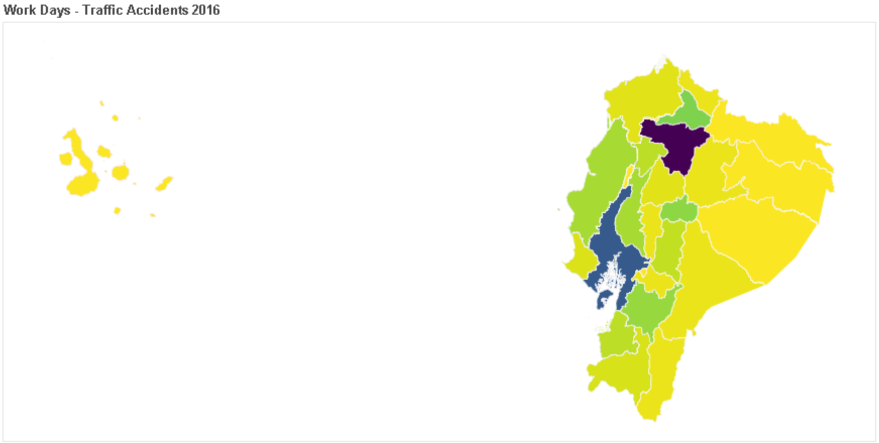

Fig. 4. Traffic accidents on work days, 2016.

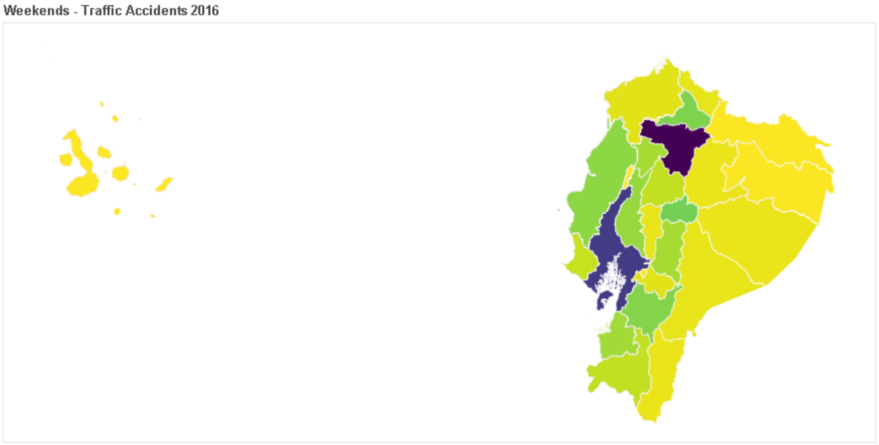

Fig. 5. Traffic accidents on weekends, 2016.

The results obtained in the first instance, did not indicate a marked variability, however, if it could be observed that in some provinces such as Pastaza, Cotopaxi and El Oro increased density, which shows a greater number of registered accidents on weekends, while in provinces such as Guayas and Pichincha the density decreases and indicates that accidents are more common on work days.

Another instant analysis that could be obtained is the contrast between traffic accidents during the seasons of the year [5-6]. Ecuador is characterized by having only two: summer and winter. These two seasons during the last years have been occurring in a very irregular way, besides considering that, the sensation is not the same at all in Coast region, Highland region or the Amazon region. The year can be divided into two parts, where the seasons converge: summer usually goes from June to December, while winter goes from January to May, although as already mentioned, this can change drastically.

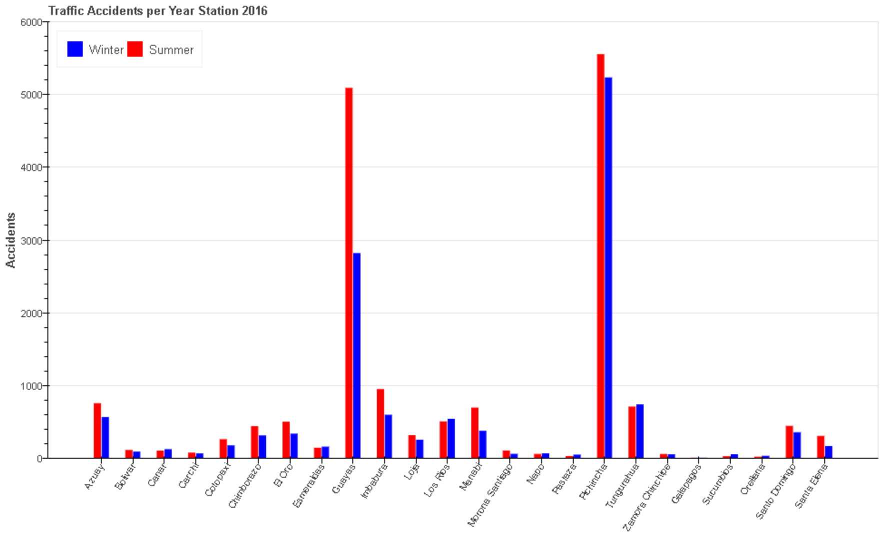

Fig. 6. Traffic accidents per weather station and province: Guayas reflecting the mean change.

Guayas undoubtedly led the position of having obtained the greatest imbalance in number of accidents with a difference of around 2000 points. Climatological variables such as precipitation [7], visibility, humidity, temperature may have influenced the increase or decrease in the probability of suffering an accident in the summer, which reaches the maximum peak.

Another way to quantify traffic accidents is according to the main class of the accident. In the 2015-2016 dataset, the accident class field classifies eight types: run over, fall of passengers [8], crashes [9], starbursts, friction, dumps, track loss and others. Taking advantage of the information of the two mentioned years, an algorithm was executed that makes a contrast between the years, in such a way that it can be inferred in which year more accidents occurred for a particular class (see Fig. 6).

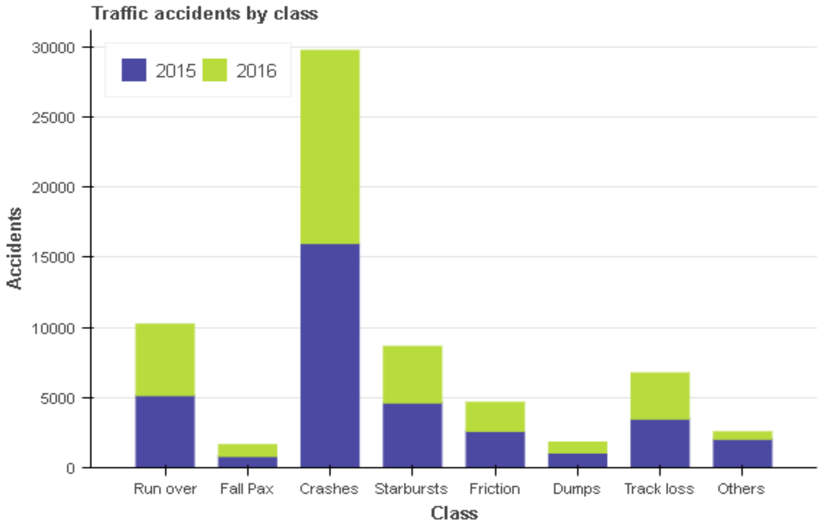

Fig. 7. Traffic accidents by class 2015-2016: the most commonly accident class is “crashes”.

The accidents for each class [10], in general, have remained constant, however, the crashes have decreased compared to the year 2015, so have the starbursts. This would indicate that global accidents decreased since also the "others" class has declined in 2016 (see Fig. 7).

The analyzes carried out up to this point have been executed with direct data from the dataset, however, starting from them, we will get to calculate other important parameters that are evaluated when a traffic accident happens, example of them, is the range of fatality in an accident.

4 Methodology

4.1 Fatality Rate or Case Fatality Rate (CFR)

One of the most important variables that is measured in traffic accident issues is the fatality rate. In summary, this measure tells how serious an accident has been. Also known as the Case Fatality Rate [11], the CFR is calculated by the proportion of the number of deaths among the total number of victims (injured and dead) in an accident. Multiplying the factor by 100, the percentage CFR is obtained. The equation that summarizes the above is defined as:

$$CFR=\frac{\#\,deathly\ victims}{\#\,total\ victims}*100 \tag{1}$$

For each traffic accident record, the respective equation of the CFR was applied. However, the fatality rate analysis was carried out by province, so that the CFR obtained will represent the average of the CFR for each registration in each province. In addition, the CFR should be comparable with other variables to determine trends, therefore the population and the total number of victims were the variables chosen to establish the contrast of trends.

4.2 Average Precipitation and Temperature per province

Using scrapping methods, I proceeded to extract average temperatures and precipitations, month by month, from each province during 2016. The purpose was to find some correlation between them or with the number of accidents.

As indicated, the data extraction was done month by month in each province and using Bokeh, a heat map was designed for both variables, which reflect average temperatures and precipitations (see Fig. 8, Fig. 9).

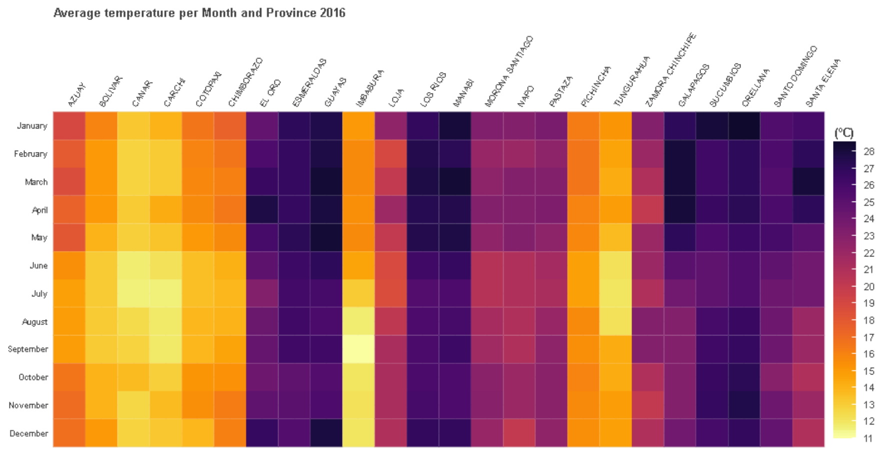

Fig. 8. Average temperatures per month and province: this graph clearly reflecting two groups well defined.

At first glance, two groups of provinces appear where the temperature ranges are well marked. In Azuay, Bolívar, Cañar, Carchi, Cotopaxi, Chimborazo, Imbabura, Pichincha and Tungurahua leads the range of 11 to 20 ºC, while in El Oro, Esmeraldas, Guayas, Loja, Los Ríos, Manabí, Morona Santiago, Napo, Pastaza, Zamora Chinchipe, Galápagos, Sucumbíos, Orellana, Santo Domingo de los Tsachilas and Santa Elena lead the range from 21 to +28 ºC.

In Figure 9, it is clearly observed that there are few provinces where the precipitations are very high and they are constant at the same time. In general, all the provinces of Amazon region are characterized for having this behavior, due to the great jungle that inhabits all this territory: Napo, Pastaza, Morona Santiago, Zamora Chinchipe, Sucumbíos and Orellana. The rest of the provinces that have the same behavior are justified by crossing the winter season, a rainy season with high precipitations.

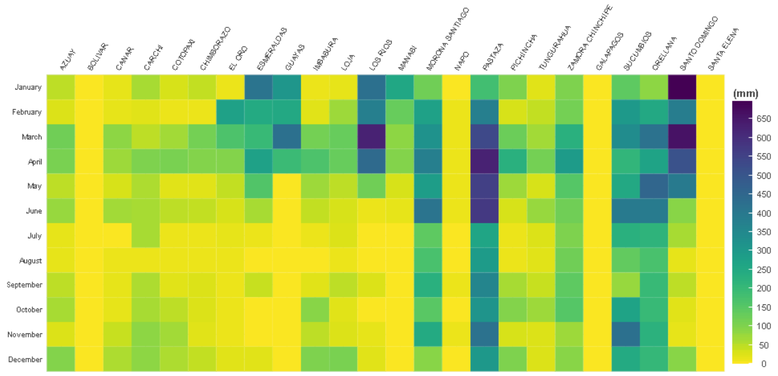

Fig. 9. Average precipitations per month and province: some of them exceed the 500 mm/month.

5 Results and discussion

5.1 Fatality Rate Analysis

As it was fixed in the methodology, the analysis consisted in contrasting three variables for each province: the number of victims, the population and the CFR (see Fig. 10).

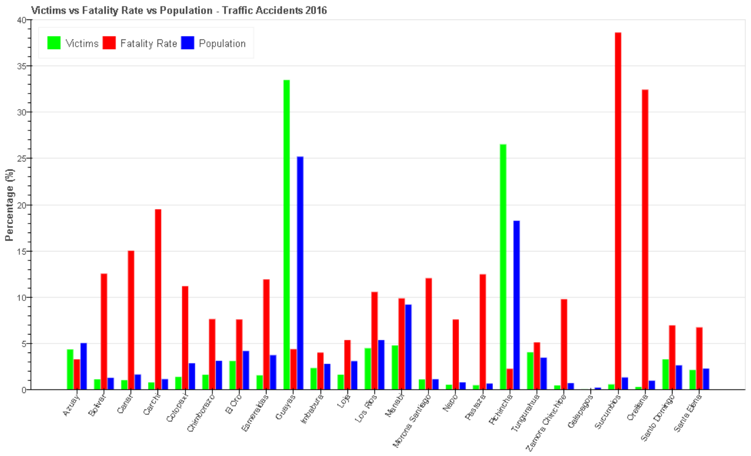

Fig. 10. Contrast between population, victims amount and CFR per province.

In most provinces, an inverse proportionality is observed between the CFR vs. the population and the number of victims. The behavior of these variables makes total sense and therefore an explanation. It must be borne in mind that the CFR measures the fatality of the accident, this is, how many people die in proportion to the total number of victims. In big provinces such as Guayas or Pichincha, the number of traffic accidents reported is quite high, and following this logic the CFR should also be, however, the inference is wrong. The graph indicates that, in a large number of accidents, such as those reported in the two provinces mentioned above, the number of deaths is at a disadvantage with the total number of victims, which is why the CFR tends to be much smaller

The same reasoning should be followed for the provinces where the number of victims and the population is low, however, the CFR is high. In this case, in the few records of accidents that occur, the number of deaths almost reaches the total number of victims, which is why the CFR is very high. Bolivar, Cañar, Carchi, Chimborazo, Sucumbíos, among other provinces, have been characterized for having this behavior in registered accidents. For the rest of provinces where the comparison is not drastic, the number of deaths is kept in balance with the number of victims, so that the CFR remains at the same level and proportion as the rest of the variables.

5.2 Average Precipitations vs. Average Temperatures

For the analysis of this point and the following two, three provinces that stand out for having the largest accident record were taken as a study reference: Guayas, Pichincha and Imbabura. Both precipitation and temperature are variable incidents in traffic accidents since they affect the environment, which largely defines the scenario of probabilities for a traffic accident. Not always these two measures are related, however, following the logic you can think that, with higher precipitation, the climate becomes cooler and therefore temperatures fall, or vice versa. The following graph shows the correlation between the mentioned variables, for the three selected provinces (see Fig. 11).

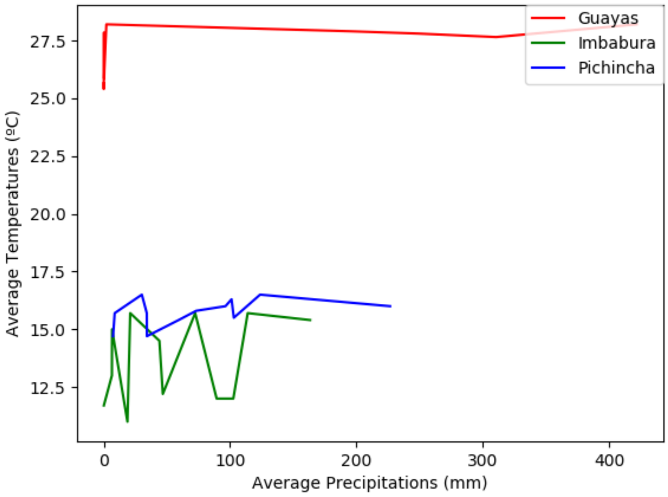

Fig. 11. Correlation between precipitations and temperatures.

Guayas maintains a constant temperature trend while precipitation increases. On the other hand, for Pichincha and Imbabura, the trend is alternating, it is not enough to describe a pattern, which denotes that there are other intrinsic variables that affect this behavior.

5.3 Average Precipitations vs. Accidents Amount

Wet roads increase the chances of suffering a traffic accident, which would indicate in an idealized model that, the higher the precipitation, the greater the number of accidents (see Fig. 12).

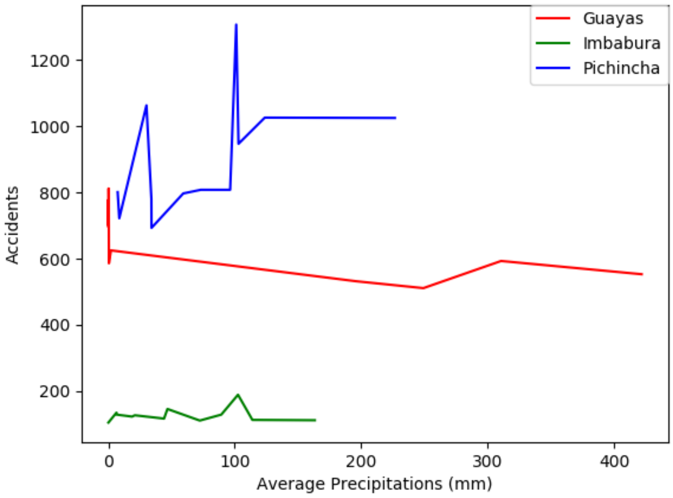

Fig. 12. Correlation between precipitations and accidents amount.

Similarly, it is observed that the graph does not follow the logic model, rather, until now it has been observed in most cases a constant trend as the number of accidents increases. Guayas and Imbabura mark this behavior, stabilizing their growth regardless of the increase in precipitations. Contrary to this, Imbabura if it is affected since varying precipitations cause an alternating tendency in the number of registered accidents.

5.4 Average Temperatures vs. Accidents Amount

It is expected that the relationship between the number of accidents and average temperatures will be even lower than the previous relationships. The temperature itself is quite independent as variable and its incidence is uncertain in most cases, however, it was taken into account for the respective analysis (see Fig. 13).

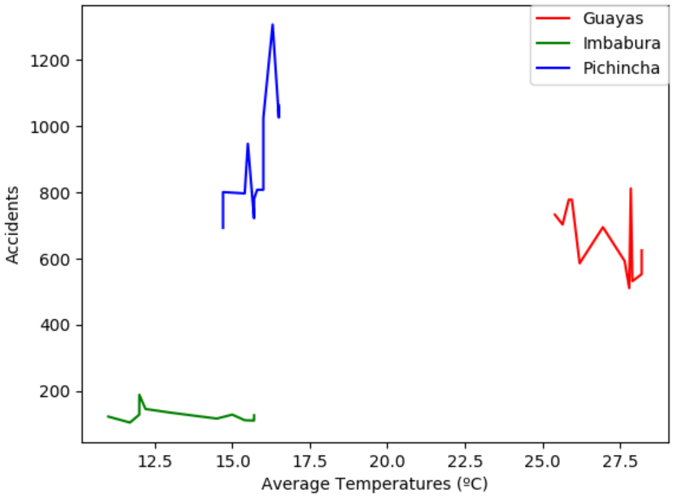

Fig. 13. Correlation between temperatures and accidents amount.

As mentioned, this relationship is even more uncertain than the previous ones. However, it can be said that the trend in Imbabura is constant, which means that temperature is not an influential factor in the number of accidents since it is always maintained at a stable level. On the other hand, Pichincha and Guayas have an alternating pattern, where it is perceived instead that the temperature causes changes in the behavior of the graph.

6 Conclusions and further Works

Bokeh has undoubtedly been a fundamental pillar in the analysis of datasets, since it has produced high-level graphs that have allowed to contrast variables and after that to establish relationships, which in certain points did not occur, but in others if they were observed trends of growth, decrease and constancy.

In general, the incidence of temperature and precipitation in traffic accidents was discovered through the analysis and study of the different graphs produced with the help of Bokeh. The institutions have a great task ahead, in continuing to dabble in the search for mechanisms that allow to enhance the reduction of accidents, one of them being prevention. The proposal of more sophisticated algorithms, methods and studies will contribute synergistically to the contribution of solutions to the different problems that exist today, in terms of transport and roads.

7 Acknowledgements

I want to express my sincere gratitude to the Web and Mobile Technologies Research Group (TAWS Club) of the “Escuela Superior Politécnica del Litoral” (ESPOL), for having been a vital support in the discovery and learning of the Bokeh library, for the analysis and study of this dataset.

References

1. Yuan, Z., Zhou, X., Yang, T., Tamerius, J., Mantilla, R.: Predicting Traffic Accidents through Heterogeneous Urban Data: A Case Study. In: Proceedings of 6th International Workshop on Urban Computing, pp. 1-9. Nova Scotia, Canada (2017).
2. Akgungor, A., Korkmaz, E.: Estimating Traffic Accidents in Turkey Using Differential Evolution Algorithm. In: Journal of Civil Engineering, vol. 12, pp. 75-84. De Gruyter Open, Turkey (2017).
3. Ali, G., Tayfour, A.: Characteristics and Prediction of Traffic Accidents Casualties In Sudan Using Statistical Modeling and Artificial Neural Networks. In: International Journal of Transportation Science and Technology, vol. 1, no. 4, pp. 305-317. Elsevier, Loughborough (2012).
4. Green, C., Heywood, J., Navarro, M.: Traffic Accidents and the London Congestion Charge. In: Journal of Public Economics, vol. 33, pp. 11-22. Elsevier, London (2016).
5. Abdelfatah, A., Al-Zaffin, M., Hijazi, W.: Trends and Causes of Traffic Accidents in Dubai. In: Journal of Civil Engineering and Architecture 9, pp. 225-231. David Publishing, United Arab Emirates (2015).
6. Narayan, S., Balakumar, S., Kumar S., Bhuvanesh, M., Hassan, A., Rajaraman, R., Padmanahan, J.: Characteristics of Fatal Road Traffic Accidents on Indian Highways. In: JP Research Center. Coimbatore, India (2006).
7. Beshah, T., Hill, S.: Mining Road Traffic Accident Data to Improve Safety: Role of Road-Related Factors on Accident Severity in Ethiopia. In: Conference Artificial Intelligence for Development, pp. 14-19. AAAI Spring Symposium, Stanford (2010).
8. Singh, H., Aggarwal, A.: Fatal Road Traffic Accidents among Young Children. In: Journal of Indian Academy of Forensic Medicine, vol. 32, no. 4, pp. 286-288. Oct, India (2010).
9. Altwaijiri, S., Quddus, M., Bristow, A.: Analyzing the Severity and Frequency of Traffic Crashes in Riyadh City Using Statistical Models. In: International Journal of Transportation Science and Technology, vol. 1, no. 4, pp. 351-364. Elsevier, Loughborough (2012).
10. Pathak, A., Desania, N., Verma, R.: Profile of Road Traffic Accidents & Head Injury in Jaipur (Rajasthan). In: Journal of Indian Academy of Forensic Medicine, vol. 30, no. 1, pp. 6-9. India (2008).
11. Savage, I.: Comparing the Fatality Risks in United States Transportation Across Modes and Over Time. In: Research in Transportation Economics: The Economics of Transportation Safety, vol. 43, no. 1, pp. 9-22. Elsevier, (2013).