Vehicle Collisions Investigated by State Police

Andra Condurache

Abstract - The aim of this study is to identify time intervals and areas that are accident-prone. The information recorded by Maryland State Police will also help to find correlations between entities involved in collisions and injuries. This investigation will focus on analysing collisions on county level. Maps are created using Tableau and some of the plots are created using Python with Spyder integrated development environment.

1. PROBLEM STATEMENT

Our life has been overwhelmed by a considerable amount of wasted time, which is unsustainable. Nowadays, the demand for mobility is increasing rapidly. One of the most serious problems we are facing is traffic congestion, which causes low productivity and a huge loss of money. More and more vehicles are becoming a part of our life. This leads to huge traffic jams. Traffic congestions can have serious consequences, on long term and obviously on short term. We will focus on short term consequences in this investigation. Collusions are immediate consequences of traffic congestions. The research questions are:

- What time intervals and regions are more accident-prone?
- Do intersections facilitate more collisions?
- Is there a correlation between objects involved in collision and injuried people?

This study will help identifying patterns that lead to vehicle collisions. One of the main aspect investigated is the correlation between the distance of the intersection and the collision, and the frequency of occurring a collusion. This may help the local authorities improve the traffic maps by introducing new routes that are going to avoid high density areas.

2. STATE OF THE ART

As extreme consumerism has became a part of our daily life, all industries have seen unprecedented increases. Our needs are totally different from how they used to be before. Over the past decades we have been producing more and more vehicles which cause serious traffic congestions or collisions in most of situations.

I have analysed 2 papers related to traffic flow and collisions. First of them is related to the relationship between traffic volume and accident frequency [2]. The authors focused on splitting accidents in categories. They calculated traffic volumes and they divided accidents into 2 groups: low traffic traffic volumes and high traffic volumes. They removed information regarding cyclists, pedestrians, wheelchairs and animals as the main interest was vehicles. Only 2336 accidents between the years 2010 and 2014 were used for this investigation. I will use a dataset with more records but spread during a shorter period of one year, 2012. Their data was from Adelaide City Council are in South Australia and my data is from Maryland State from United States of America. As these are two extremely different regions in terms of distances, average/maximum velocity, road features and lifestyle, the results may vary a lot.

The analysis used both of the spatial data and temporal data with a focus on hours intervals. I will create heatmaps to identify time intervals which are accident-prone. Retallack and Ostendorf used R programming language and RStudio integrated development enviroment. I will be using Tableau to map and visualise the data and Python with Spyder integrated development enviroment to process the data.

Wright [3] developed a model to predict traffic evolution. This study focused on the traffic flow. His model outlines potential methods of postponing traffic congestions. Numbers of lanes, maximum velocity and other rules help in predicting the critical density of the road.

3. PROPERTIES OF DATA

The dataset is provided by Maryland State Police and it is intended for public access and use[1]. It was first published on 3rd of March 2013 and it was last modified on 15th of June 2017. The dataset contains collisions investigated by the Maryland State Police in 2012 (does not include collisions investigated by local jurisdictions). The dataset contains 18639 rows and 18 columns with both temporal and spatial information.

State police recorded the exact date and time for every collision, also all the incidents are divided into 6 categories identified by codes taking into consideration the time of collision. This helps the analysis by creating time intervals and it makes easier to calculate percentages of every type of collision with the aim of identifying patterns, alongside heatmaps which contain months, days and hours.

Unfortunately, the spacial data provided was not sufficient to map the data. Another dataset containg Federal Information Processing Standards (FIPS) codes was used to correlate counties names to longitude and latitude coordinates[4]. In addition, information regarding population was attached to the initial dataset to compare collision density with population density[5]. I will focus on a column named 'DIST_FROM_INTERSECT' for studying the relation between intersections and probability of occurring a collision.

Columns 'INJURY', 'COLLISION_WITH_1', 'COLLISION_WITH_2' were analysed to answer to the third research question. Obviously some of the columns were not used during the investigation but they were analysed before computing.

Missing values

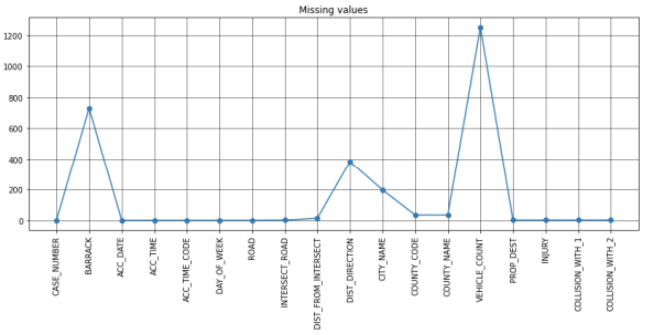

Figure 1 - Missing values

There are not many missing values. This fact facilitates the analysis. In terms of temporal features (date, time, time code and day week), there are not missing values at all. A small amount of 44 out of 18639 county names were null. These values were removed because no longitude and latitude coordinates were provided and it would have been to difficult to replace them with suitable values. An alternative way would have been to use 'City_Name' column to identify the county but almost all of the values were 'Not Applicable'. This is one of the reasons why this investigation will only focus on analysing collisions on county level as the information provided is not enough to study this topic on city level.

4. ANALYSIS

4.1 Approach

I will be using Python and with Spyder integrated development enviroment to process the data. I will be using Tableau for data mapping and data visualisation. For all the temporal analysis I will create heatmaps. Month information and weekday information will be used for the first heatmap. For the second heatmap, hours and weekday will be used. The aim of both of them is to identify high density intervals. I will identify a pattern correlated to the lifestyle. These maps are usually strong correlated to the mobility of the population. State Police ha already divided the

collisions into time interval which shows the importance of identifying time patterns. This helps the police to redesign public transport lines and redirect traffic routes where is possible.

For the spatial analysis I will need more information provided by the local authorities. I will merge information regarding population and geographical coordinates to the initial dataset. I will do this step using Spyder and I will also clean the data for each step and create another .csv file that will be loaded in Tableau.

Human reasoning will be needed to correlate information obtained from another sources. Also, I will need to removed some collisions and provide additional information. The data recorded by the State Police can not be mapped using the initial dataset.

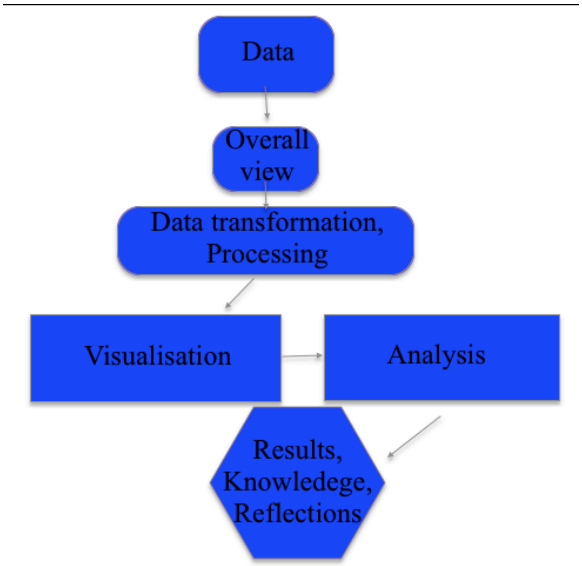

Figure 2 - workflow diagram

4.2 Temporal analysis - Weekday and Month

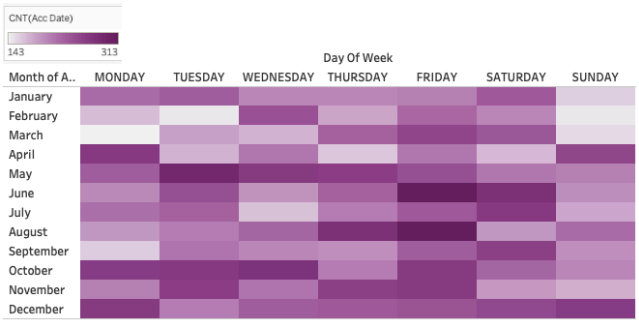

Figure 4 - Weekday and Month Heatmap

For this analysis 'Day of Week' and 'Acc Date' columns were used. The heatmap was created using CNT(AccDate) as Marks. Even though the most collisions happened on Fridays during every month, there are massive differences between months for the rest of the months.

Figure shows very low values for January, February and March, especially on Sundays. This indicates that people don't travel as much as they do during the others months. The peak values are in June, July and August, when people go on holidays. We can clearly see that the highest values were obtained on Fridays. It is an obvious behaviour because people tend to travel much more on Fridays, leaving the cities over the weekend.

4.3 Temporal analysis - Time intervals

State police divided the days into 6 times intervals coded with numbers from 1 to 6 to have a better perspective over the time of occurring the collusions.

Time code	Time interval
1	00:00 - 02:59
2	03:00 - 07:59
3	08:00 - 11:59
4	12:00 - 15:59
5	16:00 - 19:59
6	20:00 - 23:59

Table 1 - Time intervals

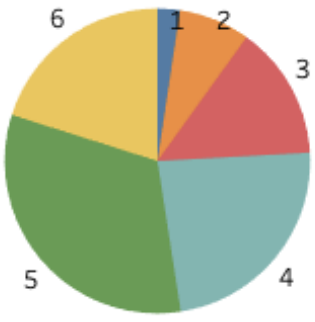

Figure 5 - Time intervals Distribution

Peak times are usually considered 6-10 am and 3-7 pm when people travel to work, but these periods may vary from city to city. State police chose different time intervals to classify the collisions which may be confusing. The lowest number of incidents occured during the night. Most of the collisions occurred during the peak hours in the evenings. This is a predictable behaviour taking into consideration that people are more tired during evenings, after work hours. There is still a considerable amount of collisions between 20:00 and 23:59. Even though the traffic volume during this time interval is definitely lower than the previous time interval, the number of collisions is amplified by various factors such as darkness or alcohol consumption.

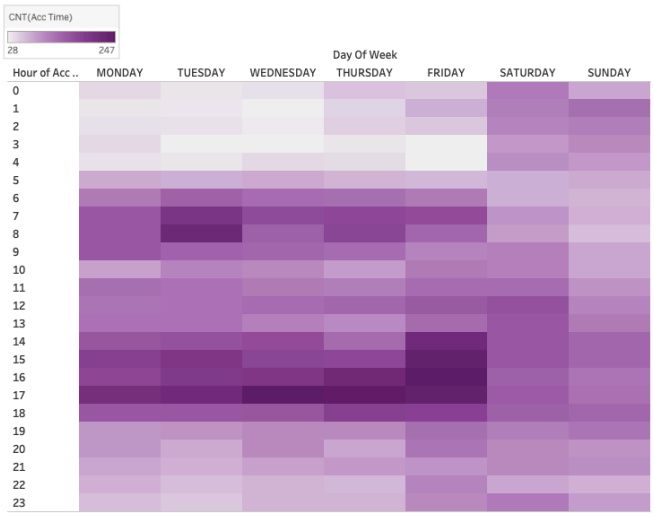

Figure 6 - Weekday and Time Heatmap

The aim of this analysis is to show off the most accident-prone time intervals. 'Day of Week' and 'Acc Date' were also used for this part but using the 'hour' component. We can observe the same pattern as in the previous analysis. Most of the collisions happened on Fridays especially between 14:00 and 19:00. In most of the cases, 17:00 - 18:00 is definitely the time interval when collisions are predisposed to occur. Figure shows that during the peak times, 15:00 - 19:00, there is a high density of collisions. This pattern is even more prominent on Friday. One interesting aspect is the comportment of first quarter interval and last quarter interval of the day. There are significant less collisions during 00:00 - 05:00 from Monday to Friday comparing to Saturday and Sunday. The exact same pattern can be seen for 20:00 - 23:00 time interval. Almost all the collisions are concentrated in a small time interval during the weekdays, while the collisions occured on Saturdays and Sundays are spread in a wider interval. This is a normal behaviour caused by people's lifestyle.

4.4 Spatial analysis - Collisions density

The initial dataset provides only county name for every collision which makes difficult to map the data as no geographical coordinates were provided.

First step was to remove the rows with unsuitable values for 'County_name' column because there were many missing values that could not have replaced with other relevant information. This process was conducted just for this analysis and the whole dataset was used for the others analysis.

According to County government, State of Maryland consists of 23 counties and 1 independent city. As Tableau is not able to map the data only based on county names, another method was used. States and counties are identified with Federal Information Processing Standards (FIPS) codes. Another dataset containing counties and FIPS codes was used to make the correlation. Every county is identified by a three-digit code and every state is identified by a two-digit code. The prefix code for State of Maryland is "24", for e.g. "24031" is the five-digit FIPS code for the county of Montgomery.

A column named "Fips" was added to the initial dataset by using merge() function. One more column "State" was added during the implementation in Tableau, having for each row the value "Maryland", with the aim of outlining State of Maryland on the map. [6]

Tableau automatically generates longitude and and latitude coordinates. "State" and "Fips" are added to Marks to draw county borders. CNT(County Name) was used to draw the map by setting it to "Color".

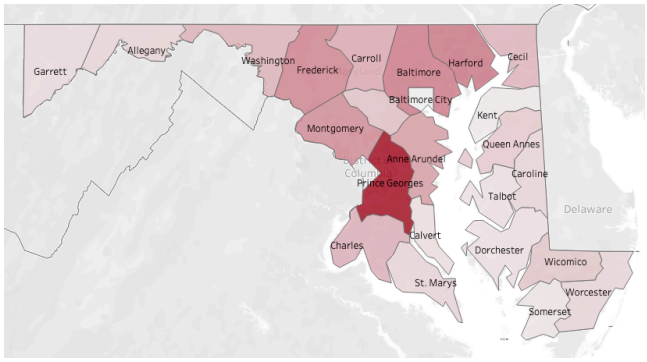

Figure 7 - Collisions Density

Figure 7 shows the density of collisions divided into counties. Prince Georges County is the area with the most collisions (3453 collisions). Montgomery, Baltimore, Frederick and Harford are also areas with very high collisions density. All types of collisions were taken into consideration for this analysis.

4.5 Spatial analysis - Population density

Same process was used to create a map to show the population density as in the previous analysis. FIPS codes were used to map the data at the county level without using latitude and longitude coordinates. In addition, information regarding county population provided by Census gov was used. I created another dataset substracting only FIPS codes for State of Maryland provisously used, county names and population values in 2012.

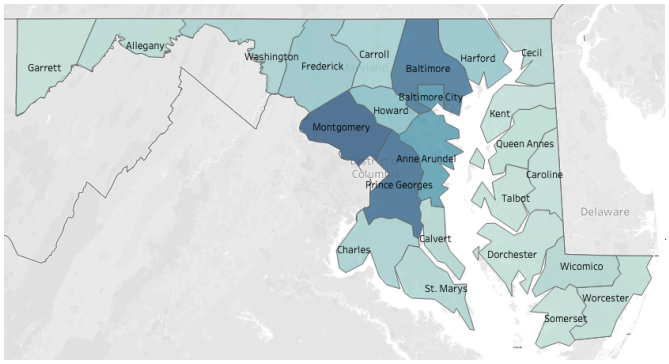

Figure 8 - Population density

The aim of the figure is to show that there is a correlation between population density and collisions density. Even though some counties do not follow the same pattern as seen in Figure 8 shows that many of the counties have a strong relation with the collisions density, especially counties like Prince Georges, Baltimore, Hartford and Fredick, which have high value in both of the analyses. On the other hand, County of Montgomery does not follow the pattern. [9]

4.6 Spatial analysis - Distance from intersection

According to National Highway Traffic Safety Administration (NHTSA), 40% of vehicle collisions occur in an intersection and this is concerning. The main reason of this result is the change of vehicle's comportment. As more than 2 roads cross each other, there are many decisions that have to be made by a driver, such as braking or accelerating, headlight flashing, change of direction or overtaking.[8] These activities lead immediately to mistakes or late reaction time. This study is focused on investigating vehicle collisions on county level so we will identify average values on county data for the first part of the analysis.

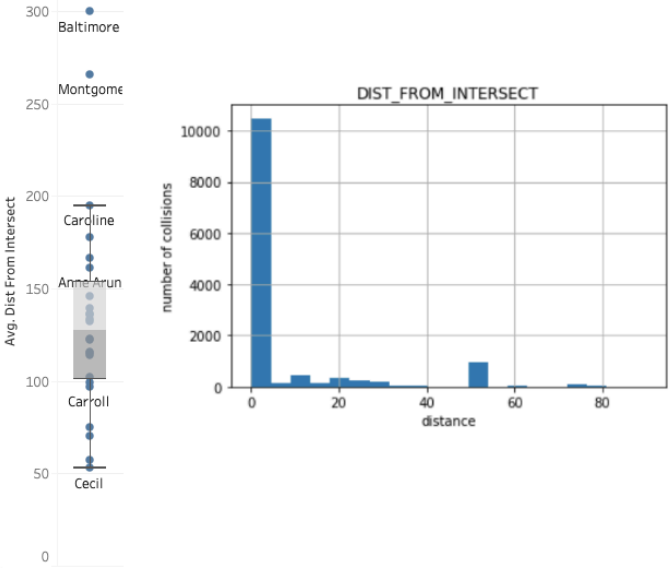

Figure 9 - Distance analysis

First boxplot was created using the average distance from intersection. Collisions were

grouped by county. For most of the counties, most of the collisions occur in the proximity of an intersection. These results may vary a lot due to geographical characteristics. Counties with very high land area have different road geometry and they have obviously more highways due distribution of cities and population. As highways are usually used for major roads, these have less intersections than local roads. Counties such as Baltimore and Montgomery have higher land areas than average they have less population than average. Consequently, there are more highways and there are more collisions occurred on highways which implies great distances from intersections. Figure shows an opposite effect for counties like Cecil, Carroll and Tablot. As the land area is smaller, collisions occurred are concentrated in smaller areas, usually located in cities, where there are more intersections.[7] This is just on overview of the average distances for collisions occurred. The second histrogram shows the distribution of the collisions occurred within a radius of 100. As I mentioned at the beginning of this part, a considerable amount of collisions occur in intersections. The safety in intersections is immensely concerning taking into consideration the results. Furthermore, the number of collisions will increase due to our lifestyle. More and more distractions are becoming a part of out life, even when driving. Careless driving is more dangerous than it used to be. Bottlenecking and lane reduction could also produce velocity fluctuations that affect drivers' reaction time which lead to accidents.

4.7 Type of Collision

I have divided all the collision into 2 categories: injuried people and not injuries people. For both of the cases, more than a half of collisions occurred with another vehicle. This is predictable taking into consideration the high number of collisions occurred in intersections.

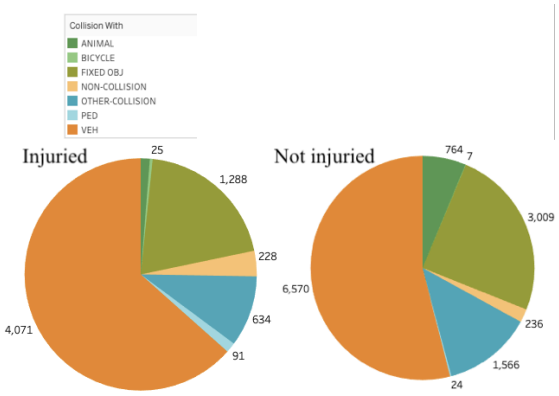

Figure 10 - Collisions with different entities

I have divided all the collision into 2 categories: injuried people and not injuries people. For both of the cases, more than a half of collisions occurred with another vehicle. This is predictable taking into consideration the high number of collisions occurred in intersections.

4.8 Results

There is a strong correlation between number of collisions and density of population which is obvious. Furthermore, our lifestyle influences a lot the probability of occurring an accident especially during peak hours. As shown in 4.6, a massive number of collisions occurred in intersection. A poor management of traffic flow leads to serious consequences as most of the accidents occurred in proximity of intersections. Collisions with another vehicles have a higher chance to produce injuries but the majority of collisions occurred with not injury.

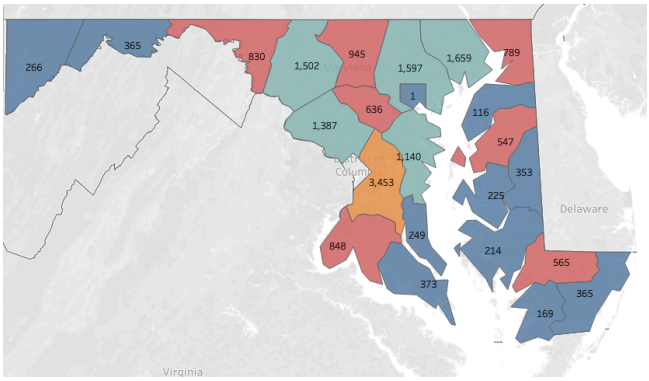

Figure 11 - Clusters

If we would cluster the data, we can definitely see that collisions are concentrated in small areas where there is a high density of population.

5. CRITICAL REFLECTION

The data used in this study was recorded in 2012 and it shows patterns correlated to our lifestyle. Creating heaptmaps helped in identifying time intervals which are accident-prone. Temporal analysis showed that commuting is probably the most important factor that leads to collisions. The most collisions occurred during the peak hours. The number of collisions will definitely decrease as a result of the recent pandemic. For late hours it will also be a massive change due to the fact that nighlife was enormously reduced. Furthermore, central crowed areas do not exist anymore as the way they used to be before.

It is still concerning that a considerable percentage of collisions occurs in intersections and in close proximity of the intersections. Local authorities should implement new schemes to divert unnecessary vehicles away from city centres. This may be possible by creating new routes with the aim of decluttering crowed areas. More detour roads should be introduced as freight transport will probably be more common in the future.

REFERENCES

[1] 2012 Vehicle Collisions Investigated by State Police, Published by opendata.maryland.gov, https://catalog.data.gov/dataset/2012-vehicle-collisions-investigated-by-state-police

[2]Wang, Pin & Chan, Ching-Yao. (2017). Vehicle Collision Prediction at Intersections based on Comparison of Minimal Distance Between Vehicles and Dynamic Thresholds. IET Intelligent Transport Systems. 11. 10.1049/iet-its.2017.0065.

[3] Wright, Paul. (2013). Investigating Traffic Flow in The Nagel-Schreckenberg Model.

[4] GIS for Maryland and Baltimore: FIPS, https://library.morgan.edu/mdgis/fips

[5] County Population Totals: 2010-2019, https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html

[6] How to Map Data on the County Level in Tableau https://medium.com/analytics-vidhya/how-to-map-data-on-the-county-level-in-tableau-9178610cd964

[7] Maryland Land Area County Rank,

http://www.usa.com/rank/maryland-state--land-area--county-rank.htm

[8] Crash Factors in Intersection-Related Crashes: An On-Scene Perspective, https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811366

[9]Retallack, A.E.; Ostendorf, B. Relationship Between Traffic Volume and Accident Frequency at Intersections. *Int. J. Environ. Res. Public Health* **2020**, *17*, 1393.

Section	Words
Problem statement	181
State of the art	304
Properties of the data	368
Analysis Approach	218
Analysis Process	1336
Analysis Results	112
Critical reflections	167