Vehicle Accidents in London:

Visual Spatiotemporal Analysis

Mohsin Gul

Abstract — To provide an aid to London's traffic management, a spatiotemporal analysis is carried out in order to analyze vehicle accidents with different factors involved that could have impact on the accidents, using the record collected by UK transport department between 2 years that ranges between 2012-2014. The spatial patterns involving vehicle accidents hotspots are presented through density maps and temporal patterns are demonstrated through bar plots and heat maps. Python is used for visualizing patterns, furthermore, guiding the analytical process. Using the DBSCAN (Density-based spatial clustering of applications with noise) algorithm clusters are identified as the areas of highest density which will be beneficial for the traffic management which can optimize and therefore mitigate the chances of vehicle accidents.

1 PROBLEM STATEMENT

A road traffic crash is an unpredictable event and can occur in various kinds of scenarios [1]. Numerous factors can be involved such as road types, road surface condition, weather conditions, pedestrian crossings etc. Between 2012-2014, 83621 vehicle accidents were recorded. To prevent these accidents and by providing safe driving instructions, data analysis is required to infer the factors impacting the accidents. This report highlights on exploring spatiotemporal dynamics of vehicle accidents in London which will certainly provide an aid to the government to raise awareness and deploy prevention mechanism. Temporal trends would help to investigate the fluctuation of crashes based on different temporal scales. Temporal question is as follows:

- What temporal trends are followed by the accidents? Is it possible to identify the specific pattern on temporal scales?

Spatial trends would help in identifying accident hotspots in boroughs in London. Following analytical questions would be addressed:

- Which areas are prone to accidents in London?
- To investigate whether the number of vehicles commuting are affected by accident prone areas?

UK Traffic Accidents dataset 2012-2014 [3] was considered for this analysis. As accidents in London were analyzed thus instances dropped to 83621 number of instances after filtering out the data. National Statistics Postcode Lookup UK dataset [4] was also used because dataset [3] did not contain region name, region code and local authority name. Therefore, these two datasets were merged on local authority (Highway) code. Region name was required as London had to be filtered out, moreover local authority names were necessary for our analysis to access the boroughs. The resultant dataset was suitable for the analysis of both types such as spatial (boroughs, coordinates) and temporal (date, time). The data possess sufficient records to carry out analysis and identify possible improvements for London traffic management.

2 STATE OF THE ART

Ganjali Khosrowshahi, Amin [2] presented the study of density based spatial clustering that recognizes the patterns accident hotspots using grid and density based GriDBSCAN, nnh (nearest neighbor hierarchical), K-means and KDE algorithms. The algorithms were applied to perform clustering to accidents recorded in Gebze and Izmit (in Turkey). The total number of 6689 observations with coordinates were used to study the accident hotspots for Kocaeli province during 2013–2014. Including several features such as existence of guardrails, curved sections, light road condition etc. For spatial analysis, density maps, bar charts were used. Out of all the algorithms employed, the author illustrates study that on comparison the most accurate method was GriDBSCAN. In my study, the data is considerably larger than Gebze and Izmit accidents data. I will not employ K-means, nnh, KDE or GriDBSCAN rather DBSCAN would be incorporated for spatial clustering to identify hotspots. GriDBSCAN is an improved version of DBSCAN by considering yielding, grid partitioning, merging with high degree similarity advantage. Moreover, I will be using heatmaps and density maps for spatial analysis.

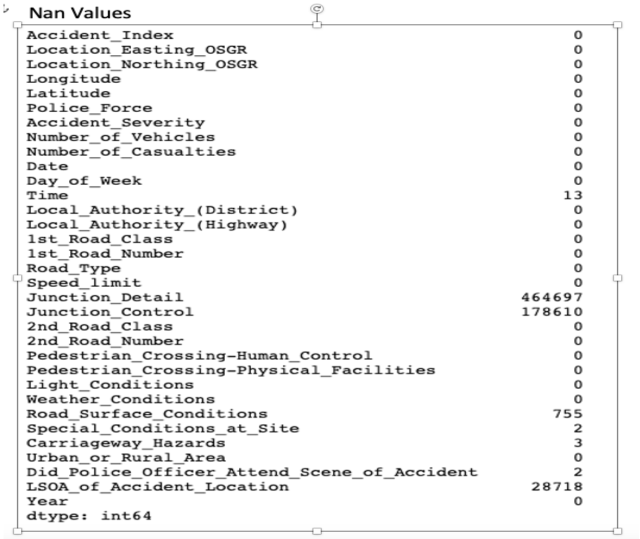
Steven Haynes [5] used the data from UK transport department of year 2006 consisting of 136621 records. The data is like our study, but the methodology of this study would be different in terms of spatial analysis as they used classification by predicting accident severity based on various factors involved such as road type, speed limit, road class and condition, light condition, weather condition, days of week and casualties. The author presents temporal analysis using line graph and bar charts. This study explores accidents and casualties on different time scales such as month, day of week and hour. The study investigates whether the time of accident occurrence affects the number of casualties. I'll be using bar charts and heatmaps to present temporal analysis.

3 PROPERTIES OF THE DATA

My data is collected by the UK transport department [3], where 287540 accidents were recorded over the period of 2 years ranging from 2012-2014, giving substantial number of instances to derive the results with the attributes used such as location, time, date and number of casualties. London is selected from all the regions; thus, 83621 observations were left. To filter out London as region National Statistics Postcode Lookup UK dataset [2] was used due to the reason that UK transport department [1] did not possess region name, region code and local authority name. The 2 datasets were then merged on local authority (Highway) code.

Temporal Analysis for this dataset is possible as date and time is given for each instance thus, analysis can be carried out with various attributes. Focusing on hour, day of week, month and year would enable me to infer the specific pattern the data is following with the accidents.

Furthermore, the data is also suitable to perform spatial analysis as coordinates such as longitude and latitude are accessible to locate the points on the map. Borough names are also provided to find the specific areas that are prone to accidents.



Nan Values

```
Accident_Index                                  0
Location_Easting_OSGR                           0
Location_Northing_OSGR                          0
Longitude                                       0
Latitude                                        0
Police_Force                                    0
Accident_Severity                               0
Number_of_Vehicles                              0
Number_of_Casualties                            0
Date                                            0
Day_of_Week                                     0
Time                                           13
Local_Authority_(District)                      0
Local_Authority_(Highway)                       0
1st_Road_Class                                  0
1st_Road_Number                                 0
Road_Type                                       0
Speed_limit                                     0
Junction_Detail                            464697
Junction_Control                           178610
2nd_Road_Class                                  0
2nd_Road_Number                                 0
Pedestrian_Crossing-Human_Control               0
Pedestrian_Crossing-Physical_Facilities         0
Light_Conditions                                0
Weather_Conditions                              0
Road_Surface_Conditions                       755
Special_Conditions_at_Site                      2
Carriageway_Hazards                             3
Urban_or_Rural_Area                             0
Did_Police_Officer_Attend_Scene_of_Accident     2
LSOA_of_Accident_Location                   28718
Year                                            0
dtype: int64
```

Figure 1– Nan Values

Missing Values

There exist missing values for the relevant features in Fig 1. Out of the 287540 observations, for 15 instances the time was missing. 755 Road surface conditions were missing. Also, we are only dealing with data that belongs to London so by filter all the values and discarding the null values we were left with 83621 observations.

Outliers

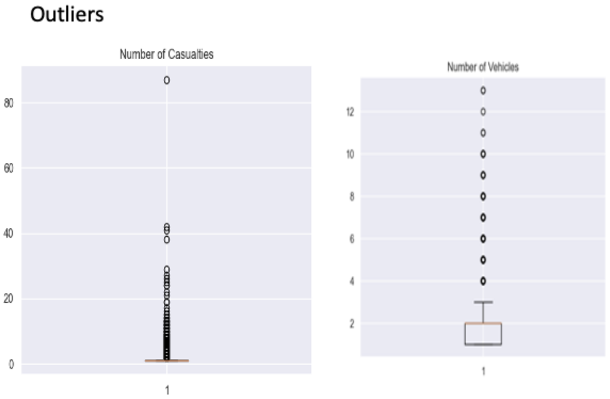


Figure 2 - Identifying outliers

To identify outliers, I plot both the number of accidents and the number of casualties in Fig 2. I used Z score method to discard the outliers. As I'm considering these two features in my analysis that is why I removed the outliers. Both Spatiotemporal analyses would consider these attributes to infer useful insights

2 ANALYSIS

2.1 Approach

I will be using Python for visualizations as it provides an ease with data manipulation, data transformation and plotting graphs. Moreover, Python would also be used for clustering where density-based clustering would be employed in order to locate the areas having maximum accidents.

For Temporal and spatial analysis, I will consider aggregation of time and coordinates with number of accidents and also casualties in a certain borough. Finding outliers before the aggregation is essential to eliminate any inconsistency in the data.

Temporal Approach – Steps

For temporal analysis, days of week, months and year would be considered such as:

1) Plot the bar plot of months across year to look if the trends are insightful in terms frequency of accidents throughout year.
2) To plot heatmap of hours with day to infer the important times of the day where there are high chances of accidents.
3) To plot heatmap of day with month to infer the important days of the week where there are high chances of accidents.

Spatial Approach – Steps

For spatial analysis, coordinates such as longitude and latitude would be considered to represent on map.

1) Identify dense areas/boroughs in London by looking at the plotted clusters
2) There is high number of instances so correct color coding would be used in order to differentiate the clusters and understand it better.
3) Plot the number of vehicles with clusters to identify if the areas affect when multiple vehicles are involved.

Spatial Approach Modelling - Steps

I will certainly be applying clustering to identify accident hotspots as it would have not been identified without clustering as visually.

1) Applying DBSCAN on data with initial parameters.
2) Now configure the values of min samples and epsilon in order to get highly dense connected clusters so that it could be differentiated from regions that are sparse
3) The clusters should be dense as when solving 3rd analytical questions, we require dense cluster in order to obtain number of vehicles that are affected by areas such as junctions.

2.2 Process

Temporal Analysis

Accidents Occurrence by Year in London Between 2012-2014

From Figure 3, it is quite evident that year 2012 had most accidents reported in these 3 years. After Dec 2012, the accidents rapidly dropped in the start of 2013 and then it gradually peaked in October 2013. At the start of 2014, accidents rose and steadily peaked in October 2014. Initial assumption is that there would be an increase in number of accidents in winters when considering the weather conditions and road surface conditions. The bar chart illustrates that for 2013 and 2014 the accidents occurred peaked in winters starting from October till January after which it rapidly drops in December. But in the case of year 2012 the bar chart fluctuates all over the year and months of March and July with the most accidents. From the span of 3 years, the bar chart depicts that the month of Feb has fewest number of accidents reported.

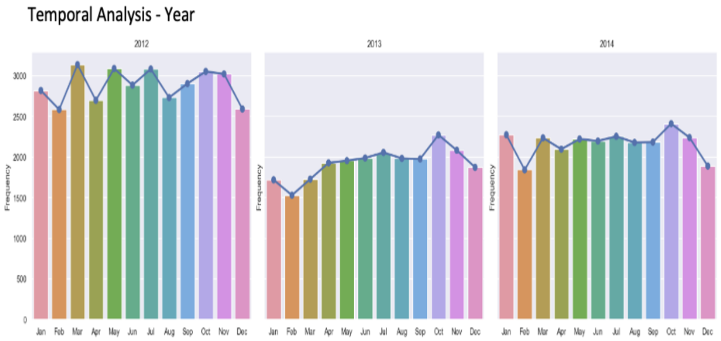


Figure 3 - Accidents Occurrence by year in London Between 2012-2014

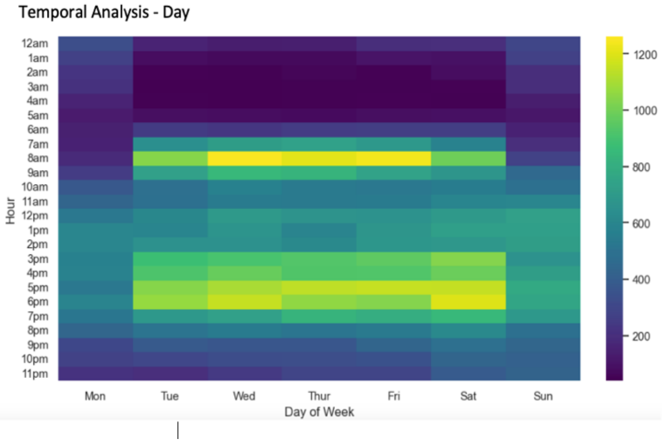


Figure 4 - Accidents Occurrence by Day in London Between 2012-2014

The heatmap in fig 4 infers that the accidents vary with time. As different time of the day shows different analysis due to the reason of dependency on people lifestyle patterns such as working, sleeping and commuting. It shows expected time of day having maximum accidents as it indicates peak times. The peak times according to London throughout Monday- Friday are during the travelling hours 7-9:30 am and 4 pm-7 pm. A slight increase in accidents 2 pm-3 pm illustrates that people go out for lunch. It is extremely surprising that Monday has fewest accidents and even in peak hours.

The maximum accidents occurred are at 8 am and in the evening between 5-6pm. This depicts that people at early in the morning and in the evening tends to reach their destination earliest as possible thus, accidents in these time frames increases.

In fig 5 the month is plotted across day of the week to analyze number of accidents happening in day of the month. In Figure 6 it is quite evident that maximum number of accidents occurred on Saturday. Furthermore, October and November have the maximum accidents occurrence.

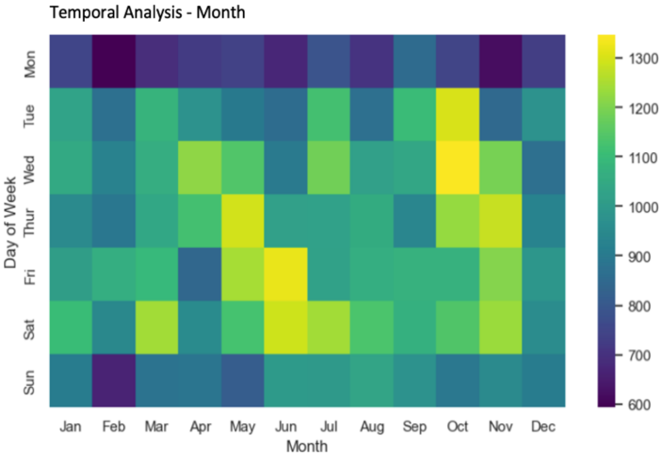


Figure 5 - Accidents Occurrence by Month in London Between 2012-2014

Spatial Analysis – Boroughs Prone to Accidents

Fig 6 shows accident hotspots with the most number accidents. The clusters formed are colored to provide a meaningful illustration. These clusters are plotted on the map without the noise. The fig 6 shows that areas such as center of London and adjacent areas, boroughs such as Tottenham, Stamford, Westminster, Wadsworth have maximum hotspots, subsequently these areas are prone to accidents, thus having higher point density than remaining boroughs.

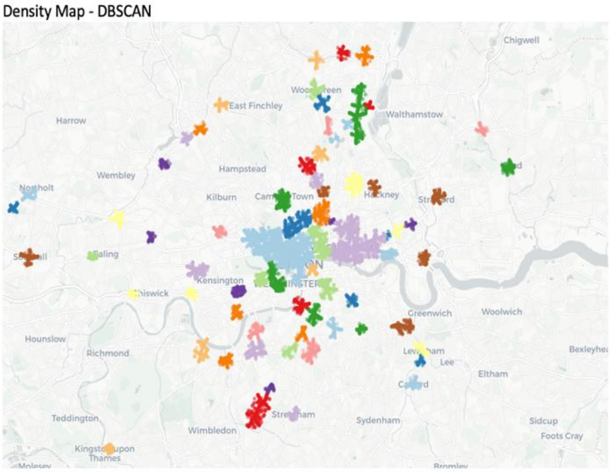


Figure 6- Density Map showing accident hotspots

Spatial Analysis – Boroughs Prone to Accidents

Fig 7 and Fig 8 infers that accidents involving two more vehicles are more likely to occur at busy junctions for instance staggered junctions, roundabouts etc. Vehicles when coming off the roundabout in Fig 9 to waterloo bridge has increased number of accidents. This could be the possibility of linkage of high-speed roads with roundabouts.

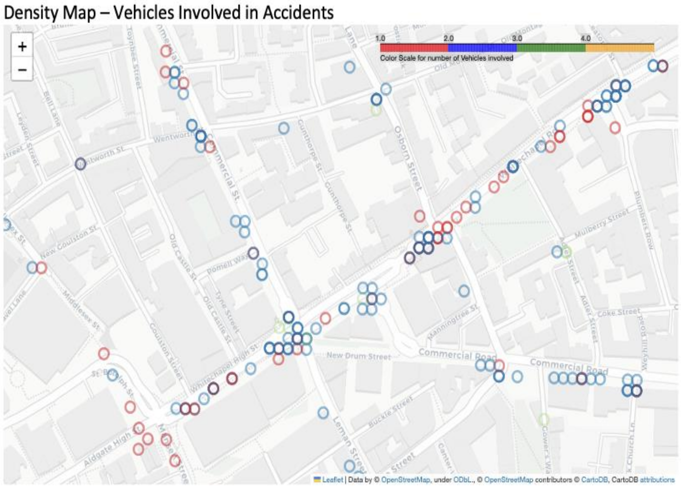

Density Map – Vehicles Involved in Accidents
1.0
2.0
3.0
4.0
Color Scale for number of Vehicles involved
Leaflet | Data by © OpenStreetMap, under ODbL, © OpenStreetMap contributors © CartoDB, CartoDB attributions

Figure 7- Density Map showing number of vehicles involved in accidents at staggered junctions

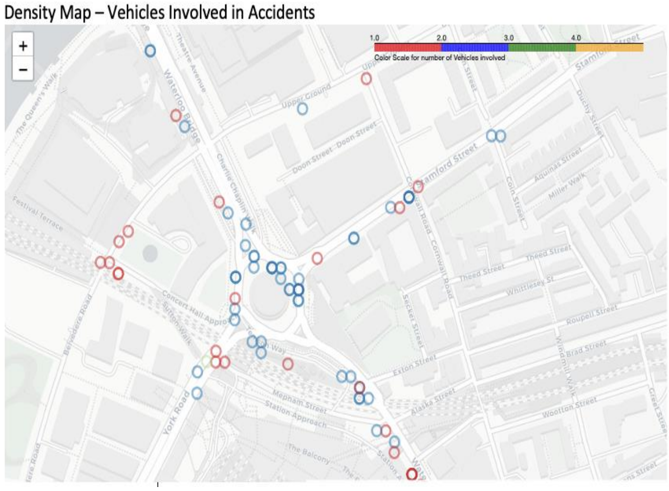

Density Map – Vehicles Involved in Accidents
1.0
2.0
3.0
4.0
Color Scale for number of Vehicles involved

Figure 8- Density Map showing number of vehicles involved in accidents at roundabouts

Moreover, the roundabout in elephant and castle interlinked with 3 adjacent roads such as London Rd, Newington causeway and St George Rd have accident hotspots involving two or more vehicles.

2.3 Results

The fig 3, fig 4, fig 5 comprising of frequency of accidents occurrence have substantial times spread. In fig 4 and fig 5 it is quite evident that Saturday has the greatest number of accidents occurred. Also, peak hour visualization of accidents will provide aid for the traffic management of London to improve the system by trying to mitigate the chances of accidents at that time of frame. Data plotted months across the year depicted a lot of fluctuations spatially in 2012. 2013 to 2014 were quite similar except of increased number of accidents in Jan and March of 2013. Fig 4 was not helpful as it fluctuated a lot but years after 2012 were insightful as it showed that Oct-Dec had highest number of accidents.

Fig 6 illustrated density-based clusters and it was successful in plotting densely connected point in a cluster and differentiating with sparse areas. It will provide an aid to traffic management to optimize due the reason that accident prone areas were discovered. Fig 7 and Fig 8 illustrates effects of areas when two or more vehicles are involved. Our findings were that areas where there were busy junctions had increased in number of accidents. This will make traffic management to plan and strategize to decrease number of accidents while considering the high-speed roads interlinked with junctions.

4 Critical reflection

Further time analysis could be carried out to find out how temporal trend has effect on the number of casualties. Moreover, the clusters could be distributed on the severity of accidents to get useful insights to help traffic management. Also, Optics Algorithm could've been used as it has property that it does not limit itself to one global parameter. In contrast, DBSCAN uses min samples and epsilon that is very difficult to choose after hit and trial method. I choose eps of 300 and min sample of 100 making dense clusters so that it is easy to differentiate the boroughs.

Further work could be done of taking speed limit into consideration and get useful insights and provide an aid to traffic management to improve their traffic signs on roads

Table of word counts

Problem statement	295
State of the art	298
Properties of the data	305
Analysis: Approach	337
Analysis: Process	499
Analysis: Results	221
Critical reflection	129

References

[1] Amiruzzaman, Md. "Prediction of traffic-violation using data mining techniques." In Proceedings of the Future Technologies Conference, pp. 283-297. Springer, Cham, 2018.

[2] Ganjali Khosrowshahi, Amin, Iman Aghayan, Mehmet Metin Kunt, and Abdoul-Ahad Choupani. "Detecting crash hotspots using grid and density-based spatial clustering." In Proceedings of the Institution of Civil Engineers-Transport, pp. 1-13. Thomas Telford Ltd, 2021.

[3] Dataset: UK traffic accidents https://www.kaggle.com/datasets/daveianhickey/2000-16-traffic-flow-england-scotland-wales?resource=download

[4] National Statistics Postcode Lookup UK dataset https://opendata.camden.gov.uk/widgets/tr8t-gqz7

[5] Haynes, Steven, Prudencia Charles Estin, Sanela Lazarevski, Mekala Soosay, and Ah-Lian Kor. "Data analytics: Factors of traffic accidents in the uk." In 2019 10th International Conference on Dependable Systems, Services and Technologies (DESSERT), pp. 120-126. IEEE, 2019.

[6] Khorshidi, Ali, Elaheh Ainy, Seyed Saeed Hashemi Nazari, and Hamid Soori. "Temporal patterns of road traffic injuries in Iran." Archives of trauma research 5, no. 2 (2016).