Analysis of UK Accident Data

Humza Khan

Abstract— UK road traffic accident analysis is performed using spatiotemporal analysis between 2013 and 2014, collated from the Department of Transport. The time trend and spatial patterns of vehicle accidents is presented throughout the United Kingdom, with further analysis focusing on Central London. Variables considered for this report accident hotspots, relationship of weather conditions with accidents, severity of accidents and number of casualties. Python Folium were used for density-based clustering and Tableau for visualizations to guide the analytical process. Area charts and heat maps were used to analyze temporal patterns and spatial patterns are conducted via symbol maps and UK density maps, to identify accident hot spots. We hope that this analysis can be used by the necessary bodies to guide their decision-making process to make UK safer.

1 Problem Statement

Temporal and Spatial analysis can be used as a guide by a variety of professionals to reduce the number and severity of accidents in the UK. Local authorities can use this data extensively to delegate more attention to areas that require work with the goal of reducing casualties that could lead to serious or fatal injuries. Road safety members can use this data to improve their accident prevention training and road safety education. Police can utilize this data to better place themselves on patrols in areas with high accidents. Weather and time data can also be studied for seasonality. The data can be used as a means of justification to introduce new laws or harsher penalties in more accident-prone areas. The temporal trends will aim to explore the questions:

- What patterns are present across time for accident frequency and severity?
- Are accidents seasonal and does this relate to number of vehicles on the road?

Spatial trends will assist to:

- Identify accident hotspots in the UK and Central London.
- Assess the impact weather has on the frequency of accidents in Central London.

The comprehensive (284,982 records) dataset used here is suitable for temporal analysis as we are provided with time and date data across two years for each of the incidents. Location co-ordinates are also provided which will help identify accident hotspots across different cities and areas. The data has important variables such as severity of the accidents and number of casualties to attribute our analysis further.

2 State of the Art

Homayoun Harirforoush, Lynda Bellalite, Goze Bertin Bénié analyzed spatiotemporal patterns of traffic accidents in urban areas of Sherbrook, Canada. They reported 7897 vehicle accidents between the period of 2011 to 2013, which was combined with detailed road network map which showed road type and speed limits. This contained 8327 segments. The data contained accident data such as date/time of collision, age and sex of drivers. The data is summarized using the severity of the accidents. Accidents were mapped with ArcGIS using their longitude and latitude coordinates. Kernal density estimation (KDE) was used to identify spatiotemporal patterns for traffic accidents across the four seasons and to identify statistically significant hot spots for dense areas. In comparison, my dataset will be using a higher number of data points, but I will be creating the same analysis in terms of spatial frequency of accidents by severity. Whilst I will not be providing analysis on road type and speed limits, I will be providing temporal analysis by accident severity, like the method in this journal and spatial analysis in Tableau and Python. Instead of using KDE, I will be using Python Folium for clustering dense areas in order to identify accident hot spots. This study is also like my study as it only looks at vehicle accidents and ignores crashes where pedestrians or cyclists are involved. This study did not use time related heat maps but used a combined spatiotemporal analysis using coordinates across time.

Khanh Giang Le, Pei Liu & Liang-Tay Lin studied spatiotemporal patterns of accident hotspots in relation to time of day and across seasons using a GIS-based statistical analysis technique in Hanoi, Vietnam. Accident data between 2015 and 2017 was used with 1132 vehicle accident records. They used a road network map which showed data such as road width, length and type. The dataset includes information such as date and time of accident, location of the accidents, vehicle types, age and gender of drivers, number of casualties etc. Bar charts are used to show the relationship between the accidents and time. The data was divided into seasons relating to Hanois weather data. Kernal Density Estimation (KDE) method was used to find accident hotspots according to time and seasons. The analysis was done with and without the severity of the accidents. The Comap method was used to understand spatiotemporal integration. I will be using a higher

number of accident points in my analysis and be using DBSCAN for density-based clustering whereas this report uses KDE to find accident hot spots. I used similar bar charts to create temporal analysis for accidents across different severity groups and time. My study will also be looking into weather conditions and seasonality across different time periods. The purpose of this article is similar to mine in the sense that it can help identify dangerous accident hot spots to traffic authorities who can then allocate relevant resources.

3 Properties of the Data

The UK accident data comes from Kaggle and this was scraped from the UK Department of Transport website that have collated data between the periods 2005 to 2014 recording a total of 1.6 million accidents. Each csv file contains three years' worth of data and each row represents a particular accident. The dataset has a total of 33 columns of information. The data was processed to remove columns that will not be used leaving us with the variables shown in Figure 1.

Column	Type	Description
Accident_Index	Ordinal	Accident unique identifier
Longitude	Quantitative	Longitude co-ordinates of the accident
Latitude	Quantitative	Latitude co-ordinates of the accident
Accident_Severity	Quantitative	Fatal, Serious and Minor
Number_of_Vehicles	Quantitative	Vehicles involved in the accident
Number_of_Casualties	Quantitative	Number of casualties from the accident
Date	Ordinal	Date of the accident
Day_of_Week	Ordinal	Day of the accident
Time	Ordinal	Time of the accident (24h format)
Weather_Conditions	Categorical	Nine weather conditions

Figure 1.

Temporal data such as date, time of each incident is given making it possible to analyze across the time of day, week and month to identify patterns of seasonality. Spatial analysis is possible as longitude and latitude co-ordinates are provided for each accident allowing me to present the accident frequency across a map. Further drilling into Central London will help identify accident hotspots. Frequency of accidents, number of vehicles involved, severity of the accidents and the resulting number of casualties will help identify and guide analysis. Weather conditions will be used to assess the correlation with the frequency and severity of accidents across time in the Central London area.

For the purpose of this report, data was narrowed to accidents ranging between the periods 2013 to 2014. 2012 data was therefore removed from the dataset. This accumulated to a total of 284,982 accidents. The number of data points support statistical significance in providing sufficient analysis using temporal and spatial analysis.

Time of the accident data was not available for 8 records and these were subsequently removed as these should not impact any trends given the large number of data points available. There were 4957 records that showed the weather conditions as 'unknown'. These were kept in our analysis to see if they had any impact on the accident frequency although further analysis will be difficult if that is the case. Other variable columns used in our analysis such as longitude, latitude, number of vehicles, number of casualties are clean and can be used fully for temporal and spatial analysis. Some variables were not used as they were irrelevant to this study or had missing values such as road surface conditions as this had missing values (481 accidents). There were also 111,259 duplicate values in the accident_index column ('2.01E+12') as identified in figure 2. However, looking at the other columns within these duplicate values showed that the data is actually individual with its own time, date and co-ordinates. I therefore decided to keep these in for my analysis.

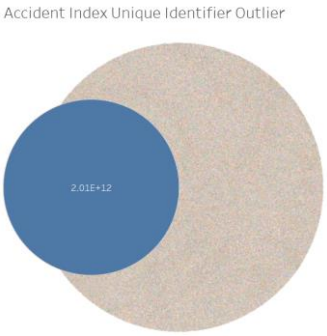

Figure 2.

4 Analysis

4.1 Approach

Tableau will be used for visualizations to create temporal and spatial graphs as this allows for manipulation and human analysis. Python folium will be used for clustering and visualization, where density-based clustering will be used to identify accident hot spots in Central London. Total number of accidents will be used but awareness must be given to individual records that can potentially skew results. To prevent this from occurring, data was checked for inconsistencies. The severity of accidents and number of casualties are used with a special focus on fatal accidents as these are what authorities should prevent as a primary goal.

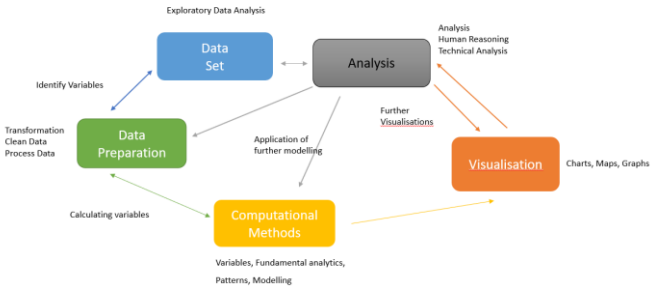

Figure 3.

For my temporal approach I will be identifying patterns of seasonality by looking at daily, weekly and monthly time data. I will be looking at total number of accidents as well by fatal accidents.

- Area graphs to show monthly accident trends across two years by severity and number of casualties to identify seasonality.
- Heatmaps to identify accident frequency by week and time of day.
- Heatmaps to identify accident frequency by week and month across two years.

For spatial analysis, I will be using longitude and latitude co-ordinates across two years of data. I will be using colour coding to identify the severity of accidents and the size of the dots to show the number of casualties.

- I will look at the accident distribution across the United Kingdom.
- Density and dot graphs across UK to identify severity of accidents and the size of the dots to show the number of casualties.
- Using dot graph to discover accident distribution by severity and casualties across Central London.
- Look at accident distribution by severity by weather conditions across Central London.

Spatial Modelling

I will be using Python folium to identify accident hotspots in Central London, which is visually better to understand.

- DBSCAN to cluster accidents within 75 meters of each other.
- I will be using Geopy in Python to consider the curvature of the earth when calculating the distance between each of the accident points.
- Identifying and plotting the clusters using Folium.
- Explaining the accident hotspots in Central London using the density-based clustering.

4.2 Process

Temporal Analysis

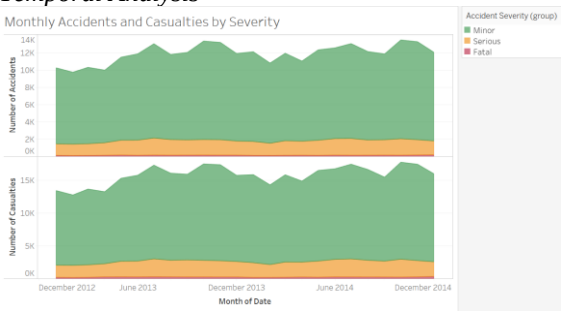

Figure 4.

As we can see from the monthly accidents figure that there has been a gradual increase in the number of accidents over the two years. The number of casualties has also increased in all three severity groups which increases the reasoning for better traffic management by authorities. There does not seem to be much variation between the number of accidents and casualties' graph which indicates that most accidents lead to some form of injury (mainly minor as this dominates the other two severity groups). The general trend observed here across both years is that the number of accidents and casualties tend to gradually increase in the first half of both years reaching highs in the summer months followed by peaks in October/November and declining in the months of December. More precisely, a seasonality pattern is observed with winter and spring having lower accidents and summer/autumn having the highest number of accidents. I would have expected more adverse weather conditions to lead to a higher number of accidents, but the data here suggests otherwise. We will explore the correlation between accidents and weather conditions for the Central London area further in our analysis.

A potentially significant goal of authorities in the UK should be to reduce the number of serious or fatal accidents. The second chart is therefore more relevant, and we can clearly see the number of fatal and serious accidents have increase drastically over the two years. From the first graph, we saw October months across both years having the highest number of accidents and casualties. When we filter away the minor accidents, the pattern here changes and the highest number of accidents and casualties are now in the months of July suggesting summer could be a factor in more fatal and severe accidents, hence proving seasonality in our data.

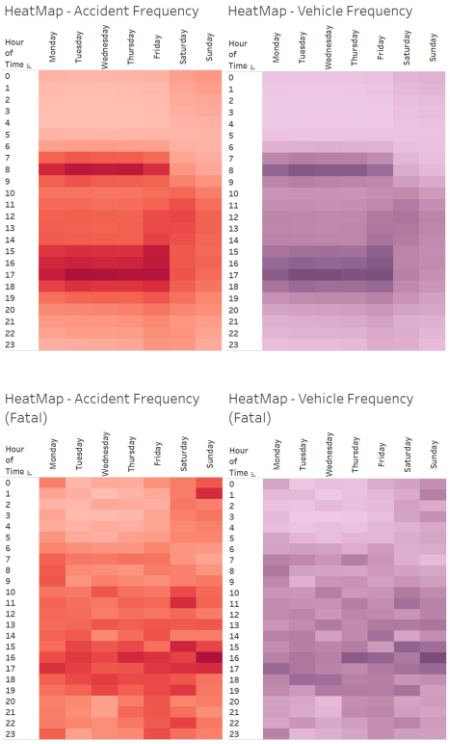

Figure 5.

Frequency of accidents is dependent on the time of day. We can see from the accident frequency heatmap that most accidents occur during peak rush hour times between 7am to 9am and 3pm to 7pm on any given weekday. This cluster of

accidents is most likely related to work travel patterns where people will be commuting to or from work or school. The number of vehicles shows an increase during the same time periods which confirms that there are more vehicles on the road during the same time periods that more accidents occur. Friday has a higher number of accidents as it is the end of the week and more people would generally tend to be out.

When we focus our analysis on only fatal accidents, we notice that the heatmap shows the weekend as having a higher number of fatal accidents. On Saturday, the most fatal accidents were at 11am and 3pm whereas Sunday's fatal accidents were recorded at 1am and 4pm. The number of vehicles on the road shows an increase in the number of vehicles during the same times. We can attribute the early hours of Sunday to possibly driving under the influence. The other times could also be travel behavior of people for example most people could be out shopping/tourism at 11am or 3pm on a Saturday.

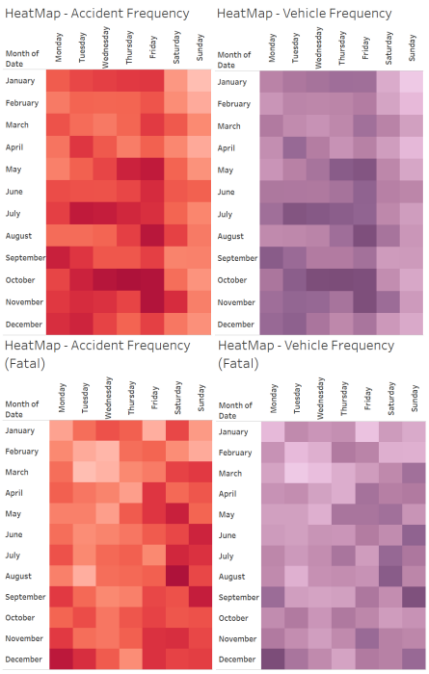

Figure 6.

I did a similar analysis with accident and vehicle frequency across each month and day of the week. Friday returned us as the day with the highest frequency of accidents although this was variable depending on the month. May, August, October and November had the highest number of accidents and vehicle frequency. The most fatal accidents occurred on a Saturday in the months of May and August. This could be due to more people and vehicles out during sunny good weather periods over the weekend.

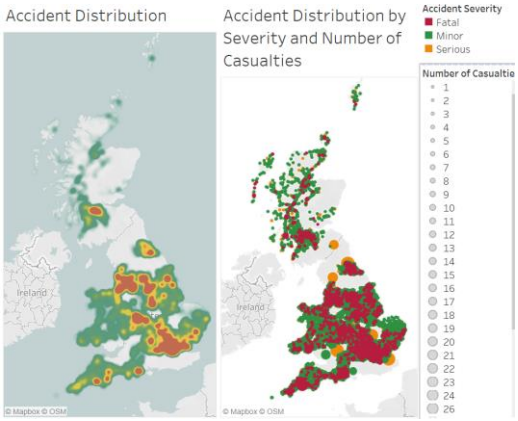

Figure 7.

The accident distribution density graph shows the accident hotspots in the UK. As we notice there is large density of accidents in major cities such as Manchester, Leeds, Nottingham, Birmingham and London. These cities are the main accident hotspots. These are main cities and more populated than other smaller parts of the UK and hence more likely that people will collide and be involved in some form of accident. The accident distribution is also shown as a dot graph color coordinated to show whether the accidents were minor, severe or fatal. The size of the dots represents the number of casualties. Whilst it is difficult to distinguish the size of dots in densely packed areas, we can notice a few big dots surrounding major cities.

Figure 8.

We focus our attention on Central London to see where the accident hotspots are. The above graph shows us all the accident hotspots which tend to be mainly minor accidents. We see a few big orange dots which shows a number of serious casualties on these roads. It is however difficult to see where the main cluster of accidents occur in Central London. We will want to try and identify the main roads where these accidents occur so it is easier for relevant authorities and bodies to focus their attention.

To be able to identify accident hotspots in Central London more clearly, I have used density-based clustering using the

Folium package in Python. This function identifies and clusters together all accidents within a given area (75 meters) and ignores accidents outside of this range as noise. Therefore, high density areas of accidents are more clearly demonstrated.

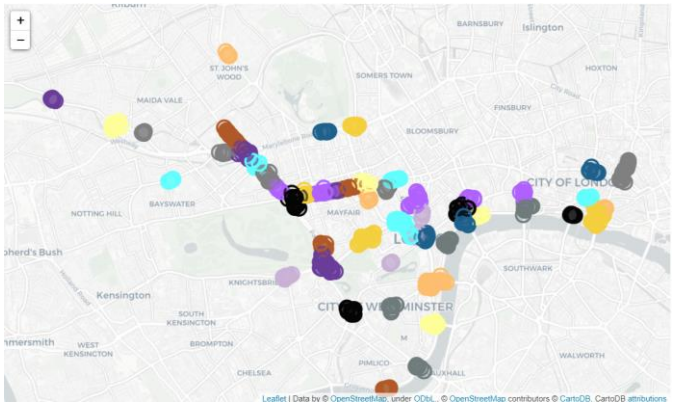

Figure 9.

As we can see the clusters gives us a better indication of where the accident hotspots are in Central London. Most of these accidents tend to occur on two very popular and well-known roads Oxford Street and Edgware Road. Oxford Street is popular shopping destinations for thousands of people and Edgware Road is a busy location filled with popular coffee shops and restaurants. Therefore, it is not surprising that these very densely populated areas are accident hotspots.

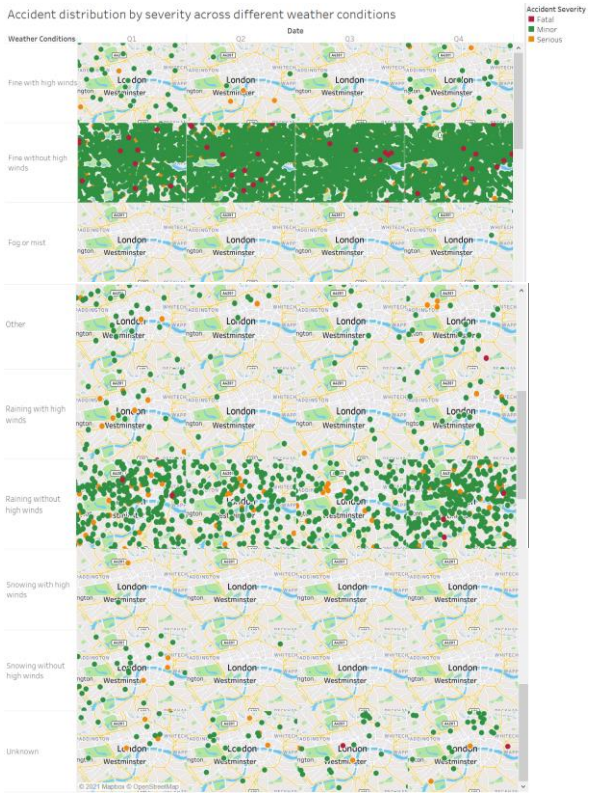

Figure 10.

The graph above shows the accident frequency by severity across different weather conditions. We notice that majority of the accidents are concentrated with fine weather without high winds. Whilst many people would assume that accidents happen during bad weather conditions, statistically looking at the above graph shows that accidents are likely to happen in good weather conditions. This may simply be because more people are out on the road enjoying nice weather, drinking, speeding etc. Adverse weather conditions are therefore not the primary reason for accidents although accidents can occur at any time. We do however note that rain can contribute to a higher proportion of accidents than any other adverse weather condition in Central London. This could be due to poor visibility or slippery road surfaces. The 'Unknown' weather condition that we kept in as discussed earlier shows that a few incidents have occurred across the two years but not being given any further detail, we cannot make any further conclusions.

4.3 Results

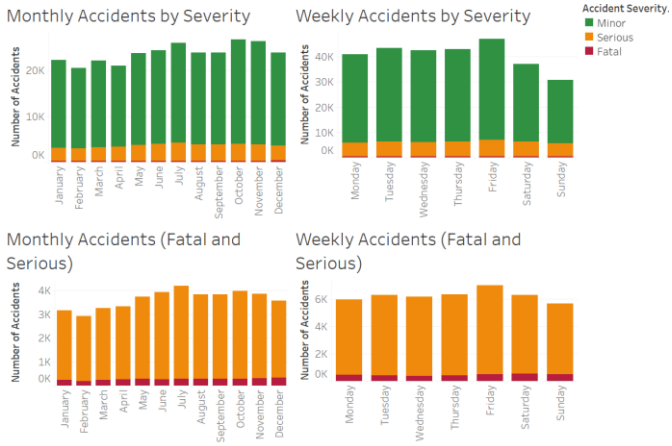

Figure 11.

We can help authorities identify the frequency of accidents by time of day, day of the week and monthly so they can strategize better to reduce number of accidents and hence casualties. We observed that accidents are seasonal and occur mainly in the summer and autumn with October across both years recording the highest frequency (see graph). The number of fatal/serious accidents show us a slightly different result with July across both years recording the highest number of accidents. We can see Friday was the most accident-prone day of the working week with more vehicles on road. The heatmap of accident frequency showed a pattern during the day which is consistent with rush hour peak travel times between 7am to 9am and 3pm to 7pm. Friday and Saturday had the most accidents when we only focused on fatal/serious accidents. The spatial density graph of the UK showed us that majority of accidents are packed closely in major cities with a higher number of casualties. Analyzing Central London showed that a lot of the accidents are concentrated along main roads and clustering together accidents showed a clearer picture which identified Oxford Circus and Edgware Road as being the most accident heavy.

5 Critical reflection

In this study we showed where and when the most accidents occurred in the UK and Central London. Minor accidents occurred more often than serious or fatal accidents. The accident distribution (Figure 3) showed that the data followed a similar pattern across both years with summer and autumn contributing to most frequent accidents. Looking at this initial graph indicated that seasonality was present in the data. I further evaluated accident frequency by month and day. The heatmap showed us that May, August, October and November had the highest frequency of accidents. We can say that accidents were seasonal, and the vehicle frequency graph showed a similar pattern showing more vehicles were on the road during the same times. Fatal/severe accidents further proved seasonality as they occurred the most in July, where we would have good weather and more cars on the road. Friday was the most accident prone on an aggregate level, however this varied according to the month. Friday was accident heavy in the summer/autumn seasons more than any other month. Higher number of accidents occurred during rush hour timings on weekdays and large cities in the UK were accident hot spots due to denser population, number of vehicles and road networks. Majority of the accidents occurred in fine weather conditions suggesting good weather to be a major factor in accidents which could be due to driver attitudes. This was not a surprising result as the initial observation showed less accidents in colder months. I was limited in the unknown weather category as I could not distinguish the weather conditions these accidents occurred in.

Further analysis can be conducted by showing the annual effects and deduce further patterns for seasonality. Density-based clustering provided me with a decent cluster of accident hot spots in Central London, but further analysis could be carried out using spaciotemporal clustering to also include regions and times most of the accidents are likely to occur. Setting the minimum sample size to 8 and finding clusters within 75 meters caused big clusters. DBSCAN concentrates on the size of the clusters and does not mention the density of the clusters. I think a Kernel density estimation (KDE) may have been more appropriate in this regard. It was therefore difficult to identify roads clearly. A tool or software which identifies roads, roundabouts, traffic light, pedestrian crossings, would provide beneficial insight into the reasons of accident hotspots. A limitation was missing data off potentially valuable variables. Demographic data on drivers involved in accidents can also provide further insight into behavioral analysis of these individuals such as age group hence better education can be provided.

There are a lot of questions that this data set can answer. Traffic volume should be considered for any future analysis because it has a correlation with accident frequency. I would suggest investigating in detail a particular accident hot spot to understand the true reasons behind that particular spot. Other spots can then be analyzed, and comparisons can be made to deduce any patterns.

Table of word counts

Problem statement	246
State of the art	481
Properties of the data	419
Analysis: Approach	350
Analysis: Process	1119
Analysis: Results	200
Critical reflection	499

References

[1] https://www.kaggle.com/daveianhickey/2000-16-traffic-flow-england-scotland-wales

[2] Qiu, C. , H.Xu, and Y.Bao . 2016. “Modified-DBSCAN Clustering for Identifying Traffic Accident Prone Locations.”

[3] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

[4] Spatial and Temporal Analysis of Seasonal Traffic Accidents.https://www.researchgate.net/publication/333601530_Spatial_and_Temporal_Analysis_of_Seasonal_Traffic_Accidents

[5] Khanh Giang Le, Pei Liu & Liang-Tay Lin (2020) Determining the road traffic accident hotspots using GIS-based temporal-spatial statistical analytic techniques in Hanoi, Vietnam, Geo-spatial Information Science

[6] N.Andrienko, G. Andirenko, G.Fuchs,A.Slingsby,C.Turkay,S.Wrobel, Visual Analytics for data Scientists.

[7] https://roadtraffic.dft.gov.uk/custom-downloads/road-accidents