Integration of Floating Car Data with accidents, Open Street Map and Weather Data for improving traffic safety

- - - - - - - - - - - - - - - - - -

Final Report

- - - - - - - - - - - - - - - - - -

Group IV: Understanding Traffic Safety

Chukwuemeka Igwe Fortune

Edwige Mukundane

David Simduwa Alsina

Janak Parajuli

17 Jul 2020

Contents

1 Introduction

Traffic safety refers to the measures that are implemented to prevent accidents on roads (Y. Maqbool, 2019). The factors that cause accidents can be categorized as: human factors, vehicular factors and roadway factors (I. Ahmed, 2013). The main users of the road are vehicles, passengers, pedestrians, cyclists etc. The ones who get mostly affected due to accidents are the main users of the roads. The approaches and efforts that are made in order to save these users from being killed or injured are called road safety. In growing Intelligent transport systems, traffic data with additional records gathered for example by police authorities, weather reports are communicated to a central database and assessed by an information system to derive traffic state conditions (T. Wang Hu, 2015) Floating Car Data (FCD) provides traffic state information in road networks and its exploitation provides potentials for traffic and safety information services. Generally there are three levels of solutions being practised for road safety. The first one is prevention of the injury and death crashes in the first place. The second is to provide 'real-time risk reduction' by alerting the vulnerable users so that they can take the mitigating measures. The third one is setting guidelines and standards for road designs, law enforcements, changing driving patterns etc[1].

This research is focussed on aiding the local police who are engaged in traffic safety by providing better understanding to the traffic accidents. The concern is on improving the traffic safety by exploring the accidents and lucky escapes with the use of past accidents, floating cars and other relevant data. In this project, we have integrated the road traffic accident data (RTAD) and Floating Car Data (FCDA) to model and predict the accident patterns in major cities of Germany. Other relevant data like open street map and weather data are also used for spatial-temporal analysis and extracting more information from floating car data.

1.1 Aims and Objectives

The main aim of the project is to analyze the Floating Car Data to enhance traffic safety. To achieve this we divided our objectives as:

1. Exploratory statistical analysis of past accidents data in terms of different variables like road condition, light condition, timestamp, place etc.

2. Integration of floating car data with past accidents and other relevant data sources like open street map, weather data etc to explore the spatial-temporal pattern of accidents.

3. Analyse the floating car data in relation to other covariates to extract the points of lucky escapes, overspeeding, black spots, hot spots and cold spots etc.

4. Modeling of the selected variables obtained from exploratory statistical analysis to predict the trend and pattern of future accidents

5. To provide python-based tools of the tasks completed in order to use them for future analysis.

[1]International Transport Forum - Joint Transport Research Centre accessed on 17th Jul, 2020

2 Methodology

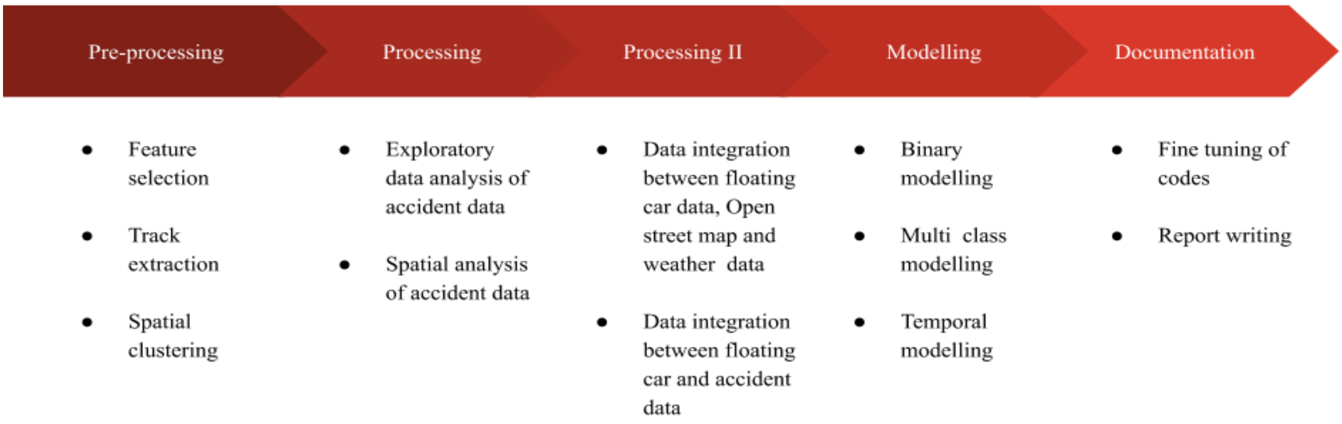


Figure 1: Workflow

2.1 Road Traffic Accidents (RTA) Pre-processing

Historical Road Traffic Data

Road traffic data of Hamburg (201-2018) and Bremen (2015-2018) were used for this research. The datasets were divided into five thematic groups: Spatiotemporal analysis, Road conditions and vehicle types, environmental variables, demographic records, and lastly the accident charecteristics.

Data Processing and Data Cleaning

The datasets were explored for the presence of outliers, consistency and accuracy. The skewness of both the datasets and flexibility in the type and metadata were also computed.

Missing/Blank value and Error Elimination

First, the datasets were checked and ranked by percentage presence of Not a Number (NAN) and not applicable (NA). For variables, a percentage greater than 40% was removed. To handle the remaining dataset with lesser percentages of missing values, since all the data were numerical and discrete, the values of kurtosis and skew (GoodData Documentation, n.d.) were measured (conclusions drawn from the histogram plotted as shown in the notebook) to determine the measure of central tendency to replace the NAN and NA numbers in each variable. Using the universal data mining principle, the NAN value was replaced with the median computed for the variable without taking cognisance of these missing values. Also, issues of bias in the subset of data were also eliminated. This process was also used for the Wrong Cell format issue confronted after deleting the text format occupying the cell. Then to remove error and outliers, each row was checked with the translated dataset chronologically and in context. This allowed for accurate and careful detection of outliers across the Bremen RTA dataset as observed in the road condition, the kind of accident, and order number variables. The binary information '0' and '1' of accident severity was created as a logical separation of the extent of accident seriousness from the "accident category" variable.

Data Selection

To enable our model to predict the type and extent of severity across the study area "Location and Severity Prediction" and "Temporal Prediction.", the following feature selection method was utilised:

1. Characteristics, Metadata and Correlation: Only those variables that were supposed to have direct relation with accidents were selected based on the literatures from (Bülbül et al., 2017; Ding et al., 2010; Thomas, 1996). For instance, "participant date of the birth month", "issue license data of birth month," and "street name" variables only complicate the data and have no relevance to the main objective to be achieved; hence, they were removed. This method was used concurrently with the correlation matrix between the different variables, to ensure none is a replication of the other. Variables that shared a strong correlation (multicollinearity) between each other were assumed to be duplicate and were removed.

2. Using Parametric method Test independent method of the chi-square class (χ^2) was used to compare the observed and expected values using 'accident severity' as dependent variable. The statistical significance of all variables were also checked at a probability value of 0.05. In addition, linear regression with 'accident severity' considered as outcome was derived using scikit learn linear model taking the probability (p-value) with level of significance (α). If the p-value is above 0.05, we delete the function; otherwise, we consider it.

3. The methods above were checked using the Recursive Feature Elimination (RFE), which recursively removes attributes while reiterating the process and building a model on those attributes that remain (Chen et al., 2018).

2.2 Floating Car Data Processing

Data integration and Variable selection

Floating car data (FCD) was acquired using the EnviroCar API for a defined bounding box. Open street Map (OSM) was acquired via the OSMNX library. Weather data was also obtained from online weather sites. The projection system of FCD and OSM were made uniform and the date and hour from weather and FCD werre compared. For our purpose, we only selected Trackid, geometry, speed, and the time stamp from FCD. Only speed limits and Junction locations were extracted form OSM. From the weather data, we extracted precipitation, visibility and wind speed measurements as explained in (Chrzan Smal, 2015).

Mapping flags and Spatial Clustering

Variables values were given flags of 0 or 1 depending on whether their categorisation is perceived as leading to high or low accident risk. Scaling the variables into numerical flags facilitated subsequent clustering for data labelling. We opted for an unsupervised learning approach, using Expectation Maximization (EM) method of K-means clustering technique. The EM algorithm computed classification probabilities to estimate data value distributions that are based on mixtures of varied distributions in other clusters (Lücke Forster, 2019). At the end data was assigned to a given cluster based on the highest classification probability. We created two sets of clusters: one for low likelihood (risk) accident instance and the other for high likelihood accident instance. Each track event was given a cluster label and mapped as low or high accident risk.

2.3 Analysis of RTA and FCD

Statistical Analysis of Accidents

This was done as a precursor to enable the modelling and support the analysis of the Spatio-temporal and semantic patterns of the floating car data (FCD). The visualization of data was explored thoroughly, and various conclusions were derived using the plotly.graph_objects, plotly.express and pandas library.

Exploratory Spatial Data Analysis

To check the spatial auto-correlation of the variables of interest, dot map and spatial clustering were prepared. The dot map was prepared to analyse the existence of spatial patterns in the distribution of the variables classified into three classes using natural breaks (Jenks) classifier. No normalisation was done at this stage, as the objective was to observe the spatial behaviour of the chosen variables independently.

For spatial clustering, we produced a heat heap to visualise accident intensity. Then we then employed a density-based clustering (DBSCAN) technique to identify accident hotspots in Hamburg for the year 2016. It clusters those points that are closely grouped together and categorizes those points that are outside of the clusters as noise (Daszykowski Walczak, 2009).

Points that are not assigned to a cluster were given the label -1. DBSCAN utilizes a metric when calculating the distance between points. To account for this, we created functions that use the latitude and longitude of two points and compute the distance in between them in meters. The great_circle module from the geopy library was used to guarantee that the curvature of the Earth is preserved.

Analysis of Lucky Escapes and Overspeeding points

Speed value was extracted from the floating car data from the envirocar API to explore the over-speeding points by comparing against the recommended road speed limit extracted from the OpenStreetMap data. The accident data wads used to generate hotspots and cold spots for accidents using the GetisOrd function, a very popular and commonly used statistical method for calculating clusters. The accident hotspots and over speeding points helped in identifying those lucky escape points in which an accident could have occurred but did not. Data for bars and pubs for the study area were obtained from the OpenStreetMap nominatim API. Location of accidents within a 1km buffer to a bar was considered to see if they played some significant role in road traffic accidents.

A function was devised to compute the difference in speeds of the current point from that of previous to find out the deceleration. For the roads which had more than one speed limit values, the lowest one was considered and for those with no limits a default value of 130kmph was assigned. To investigate the points that had a sudden reduction in speed, OpenStreetMap (OSM) data was then acquired and spatially joined using a spatial join function from the geopandas python library based on the geometry of the points of speed reduction from the floating car data from the enviro car API.

2.4 Modeling and Prediction

Both datasets were used for designing three different models for predicting the probability of accidents, classification of accidents and trends of accidents. Two Artificial Neural Network (Binary and Multiclass) models were designed and one Recurent Neural Network (RNN) was designed for temporal analysis. The datasets were preprocessed and scaled before using in the models. The datasets were splitted into train and test sets before designing the architecture of the data to segregate the relations within the data. Using the train set, it was compiled and fitted into the designed model and prediction was made with the test set. To quantify the trust-worthiness of the model, it was evaluated using different statistical measures including mean and variances also. To prepare the dataset for RNN, we had to customise the Euska data for the city of Bremen to extract the time and number of accidents information. It was then stored in a Comma Separated Value (CSV) format which was splitted into two CSVs one for training the model and another for testing.

3 Results and Discussion

3.1 Preprocessing

After data cleaning, correlation matrix was prepared to find out the variables in Hamburg Data showed almost no relation. This was opposite in case of Bremen Data, as several strong correlations were found and removed.

From the above correlation matrix above, multicollinearity was checked and the following discovered:

```
operation_number =(similar to) month
month = season
day_of_year = month
hour_cat = hour
kind_of_accident = type_of_accident
internal_information_number_2 = feature_y
participant_internal_information_number_2 = feature_y
longitude = feature_y
latitude = feature_y
feature_x = feature_y
participant_date_of_birth_year_2_new = participant_issue_date_of_the_driving_license_year_2_new
date_of_birth_year_new = issue_date_of_the_driving_license_year_new
```

> But for visualization, the `month` and `season` variables are not removed as well as the `coordinates` in the two reference systems...

Figure 2: Reducing collinearity among the variables

As expected, variables within the same thematic group, such as demographics and Spatio-temporal characteristics, presented high correlations between themselves. However, some associations between thematic groups were also detected; for example, a positive correlation between operation number and month of accident was seen. Further investigations were done using parametric and the RFE methods. The result of the ordinary least square (OLS) was positive when compared to that of the RFE method. Hence, the OLS result was utilized. In the figure below, one can see about three selected features for Bremen and Hamburg data using this method shared similar characteristics.

```
OLS(df_num, 'acccategory')

['kindocollision',
 'typeoaccident',
 'lightingcond',
 'passengercar',
 'pedestrian',
 'motorcycle/scooter',
 'hgv']
```

(a) Hamburg

```
OLS(df2, 'accident_categore')

['index',
 'day',
 'month',
 'year',
 'minute',
 'number_of_participants',
 'type_of_accident',
 'road_conditions',
 'location_innercity_extraurban',
 'street_class',
 'kind_of_accident',
 'gender',
 'amount_of_property_damage',
 'amount_of_property_damage.1',
 'longitude',
 'feature_x',
 'day_of_year',
 'casualty_age',
 'casualty_gender',
 'experience',
 'accident_severity']
```

(b) Bremen

Figure 3: Feature selection of the cities

3.2 Statistical Analysis of Accidents

The obtained results when statistical analysis of accidents was done can be summarised as follows:

1. In both the Hamburg and Bremen data for all the years accounted for, January had the least number of accidents, and November had the greatest number of accidents. The most likely explanation for this can be due to the severe cold at this time in Germany. However, it changed a bit after we broke down the overall accident count by severity in two, i.e., a grouped bar chart that the month of July had the least number of accidents for Severity type ”Not Serious” accidents, while February had the least severity classified as ”Serious” for Bremen. For Hamburg, January remained the least amount of accident for both classes of severity.

2. Autumn Friday was found to be having the greatest number of casualties for both data while Sunday (Summer) had the least number of casualties for Bremen. For Hamburg, Suymmer Wednesdays had the highest number of severities, and the least number of observations found to be Sundays of Winter.

3. For the time of day analysis in Bremen data, we found that for weekdays the number of casualties peaked during the typical commute hours: 8 AM and 5 PM each day. For weekends, the trends were different from weekdays as we see a less defined peak in the number of casualties at around noon. After 12 PM, the number of casualties slowly decreased for the weekend days. This data was not available for Hamburg RTA.

4. When rural and urban accidents in Bremen were analysed, we found that about 80% of casualties in were in rural areas. We also found that most urban casualties were severe, with a severity rating of ”Serious.” Compared to urban areas, rural areas had many traffic accidents with the least severity. Rural accidents with the severity of ”Not Serious” made up 87% of total casualties, and rural accidents with a severity rating of ”Serious” (the most severe) made up 12% of Serious overall severity. When compared to rural areas, urban areas had a similar number of accidents with severity in reverse being responsible for more ”Serious” severity rating. With the Urban and Rural breakout, we also looked at casualties by road type and determined that the majority of traffic accidents occurred on a community road type for street class.

5. Higher percentages of severe accident severity involved longitudinal traffic and collision with crossing vehicles which relate to cross junction. This may relate to the absence of traffic lights around those locations in Bremen. For Hamburg, longitudinal traffic was responsible for the highest severity of “Not Serious” casualty while turnaround crossing was responsible for the highest “Serious” injury.

6. For Bremen, men were more responsible for ”Not serious” accident severity, while women were responsible for ”Serious” accident severity. The age of 20, 45, and 55 years for males were more accountable for Not Serious accident severity and accidents in general.

7. For data of both places, dry road conditions had more correlation with the number of accidents, and the daylight conditions had more presence of accidents. This can be inferred due to a more extended dry period throughout the year and that activities by commuters are carried out during daylight.

8. Overall in Bremen, Autumn had the highest number of accident severity. For Hamburg 2016, Summer had the highest accident severity, which was not the case for Bremen.

9. Finally, 2018 Bremen had the highest increase in accidents between 2015 and 2018 with a drop in 2016.

3.3 Exploratory Spatial Data Analysis

Spatial Analysis

The figure 4 shows the distribution most affected communes of Bremen. At a glance, we can see that the age between 38 and 51 years old are more responsible for accidents.

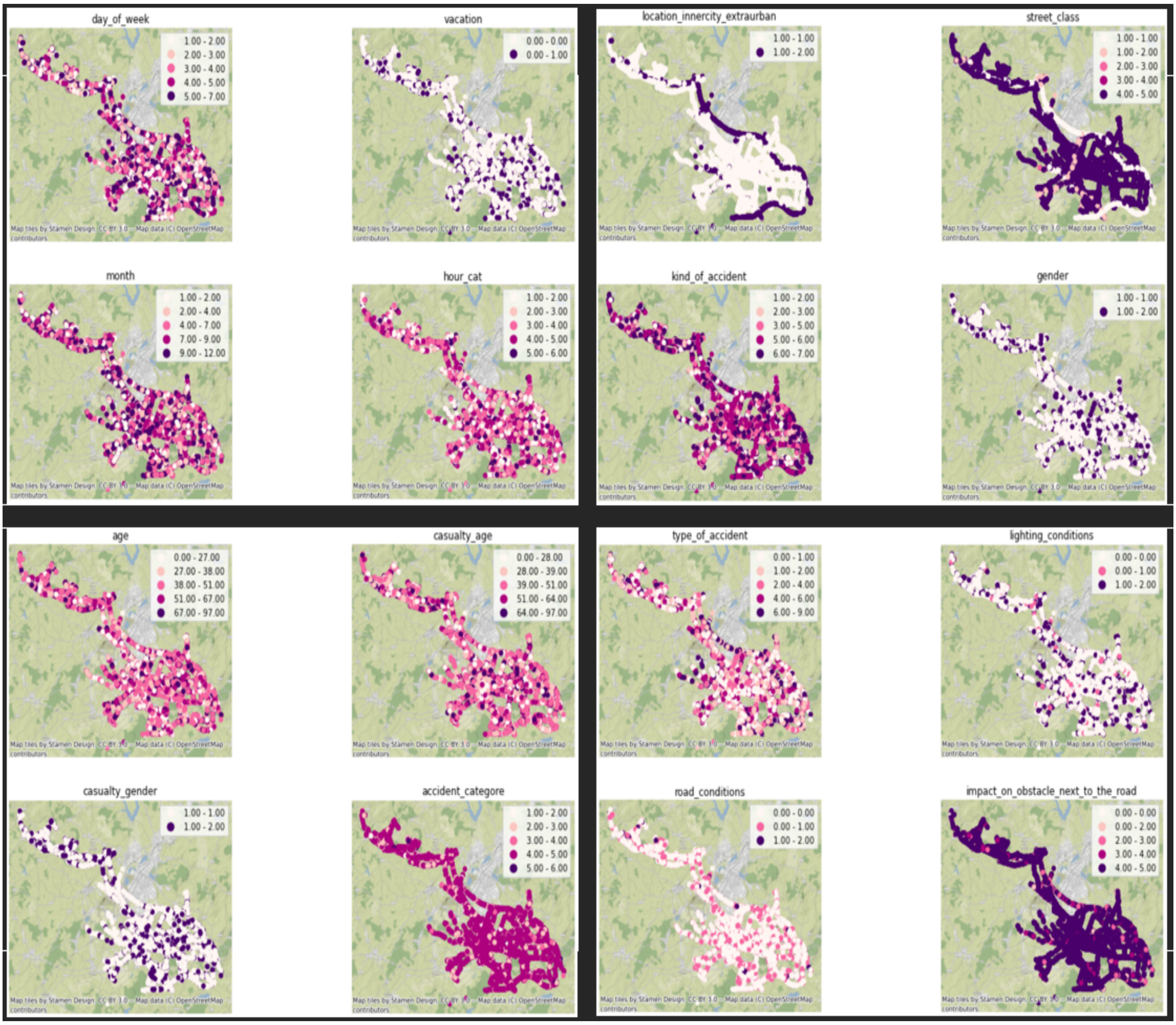


Figure 4: Spatial Data Analysi

It also shows that there are more casualties of male on dry road conditions. Also more urban road traffic accidents especially around the city centre of Bremen can be seen. Several variables tend to vary in value across the study area (type_of_accident, kind_of_accident, gender, day_of_week, vacation, month, hour_cat, and casualty_gender) while others have a spatial trend that can be considered as a cluster (age, casualty_age, accident category, street class, location_innercity_extra-urban, impact on obstacle next to the road).

Spatio-Temporal Clustering and Hotspot Regions

The spatial clusters prepared using folium gave the similar result of accidents clusters concentrated on the junctions of the roadways. When it was visualized in a space time cube, it was found that most of the crashes were seen between the duration of 07:00am through 08:00pm.

Using the haversine metric, the clustering of accident prone areas was computed and represented using a leaflet map. The figure below shows the road's vicinity with a high risk of accidents.

(a) Spatial Clustering

(b) Space time cube representation

Figure 5: Spatio-Temporal clustering of the data

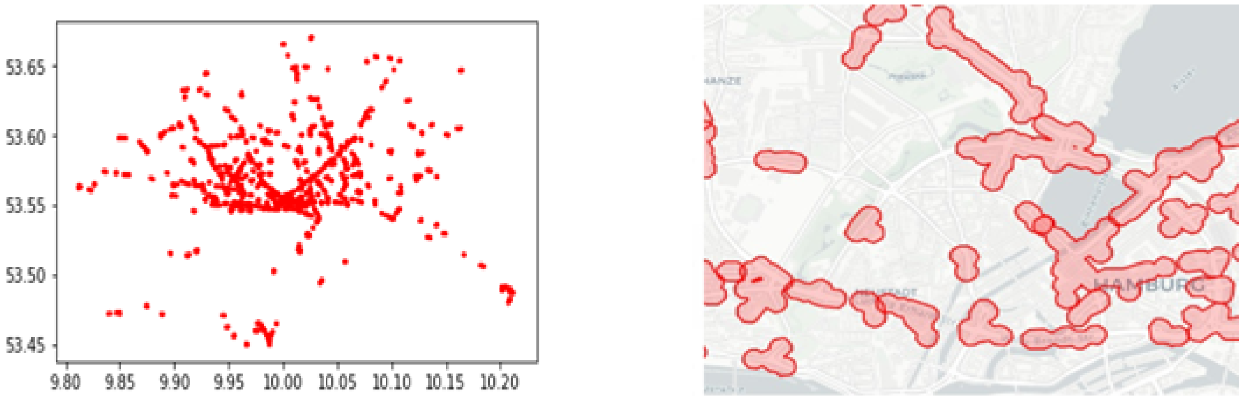


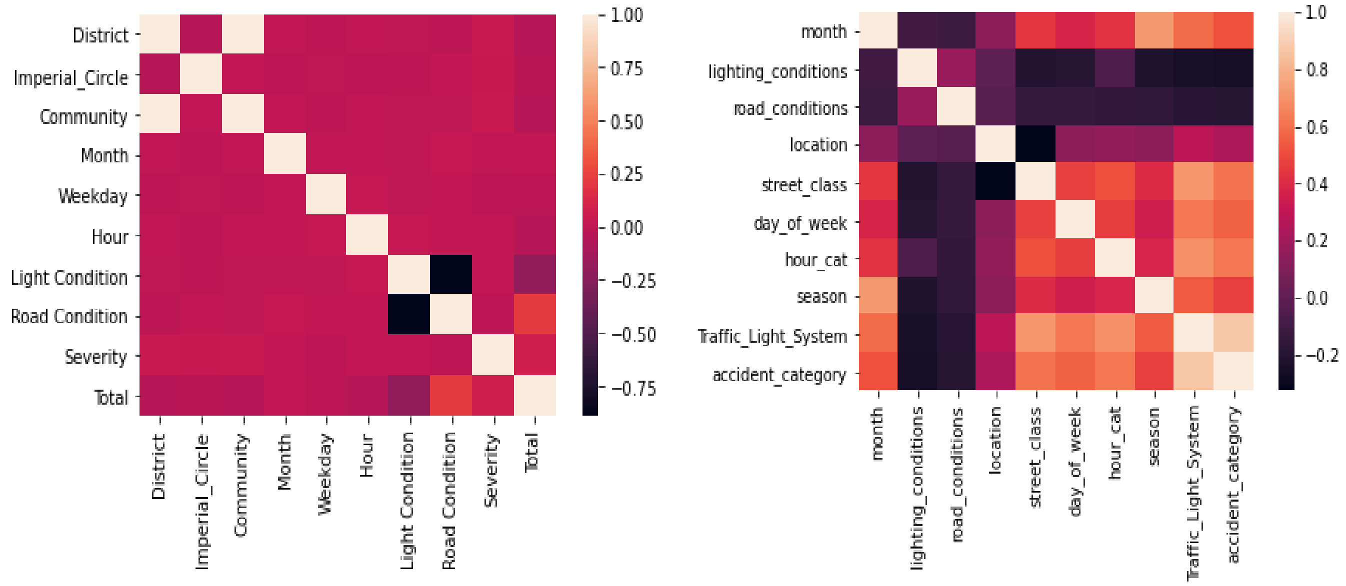
Figure 6: Hotspot regions produced from Haversine metric

3.4 Investigating Spatial Autocorrelation

Variable	Moran's I Statistic	P-value
type_of_accident	0.10803	<0.001
kind_of_collision	0.07624	<0.001
month	0.00425	<0.002
vacation	0.00356	<0.029
road_conditions	0.01607	<0.001
lighting_conditions	0.01846	<0.001
hour_cat	0.01609	<0.001
age	0.02322	<0.001
casualty_age	0.01227	<0.001
gender	0.01430	<0.001
casualty_gender	0.02285	<0.001
impact_on_obstacle_next_to_the_road	0.06371	<0.001
location_innercity_extraurban	0.72409	<0.001
street_class	0.79434	<0.001
day_of_week	0.00366	<0.018
accident_categore	0.03801	<0.001

Figure 7: Global spatial correlation test of the selected variables for Bremen

All variables included in our analysis appeared to present strong spatial autocorrelation, according to Moran's I statistics. Therefore, we can expect that the model, although not considering the geographic space in the network, will present clusters that are relatively clustered in space. After all this, a final selection of variables were made for Hamburg and Bremen data since they were found to be sufficient to design our model: **Probability of severe accident, Classification of accidents and Future trends of accident.**



(a) Hamburg (b) Bremen

Figure 8: Correlation matrices for variables of cities

From the correlation diagram of the final selected variables of Hamburg RTA data (left), it can be seen that accident severity has similar relations with all the selected variables. It is a bit positively related to road condition and negatively to light condition. From the correlation diagram of the final selected variables of Bremen RTA data (right), it can be seen that the accident category has intense negative relations with road and light conditions and positive with the traffic light system.

3.5 Floating Car Data Analysis - Integrating FCD and other Data

Accident event with Floating Car Accident Risk

By using the accident data of 2018 of Hamburg, an accident heat map was prepared to depict vehicle movement in relation to high intensity accident areas. This map aid in depicting places where accidents can re-occur.

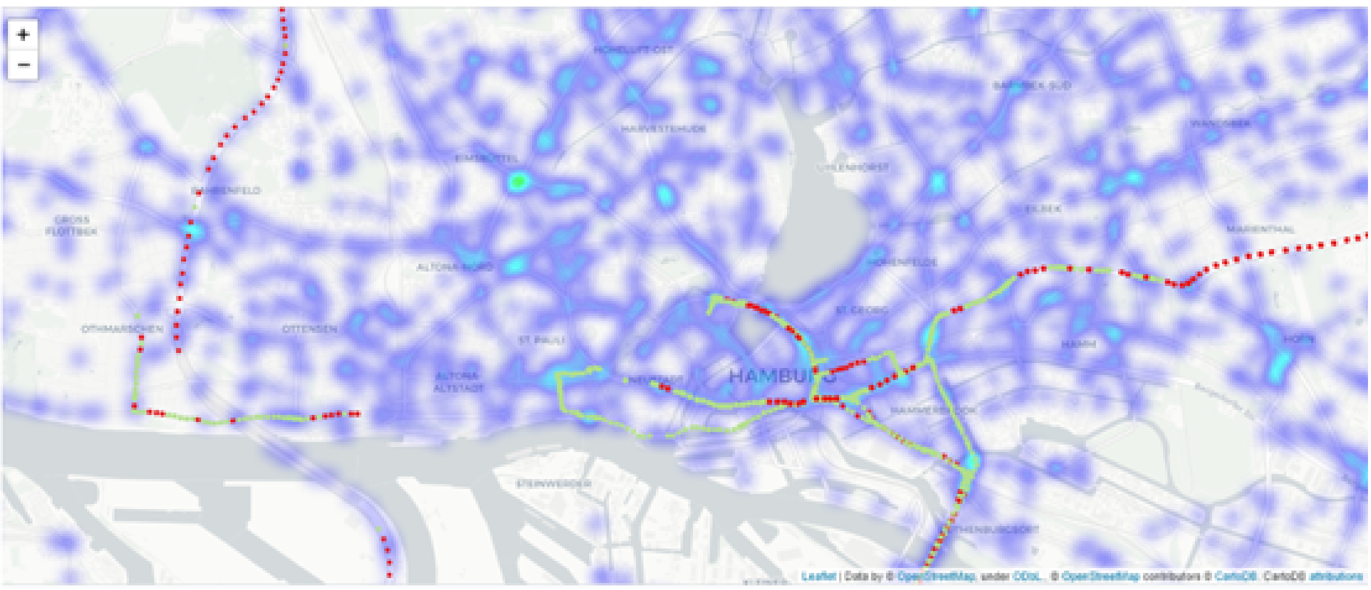


Figure 9: Heatmap of accidents

Temporal analysis

When the accident risk prediction was plotted as a violin plot in temporal dimension as in figure below, it was found that any part of the day between 07:00am through 08:00pm can experience peak accidents than other parts of the day. Also it is recommended to improve the driver behavior with consideration to the then weather condition.

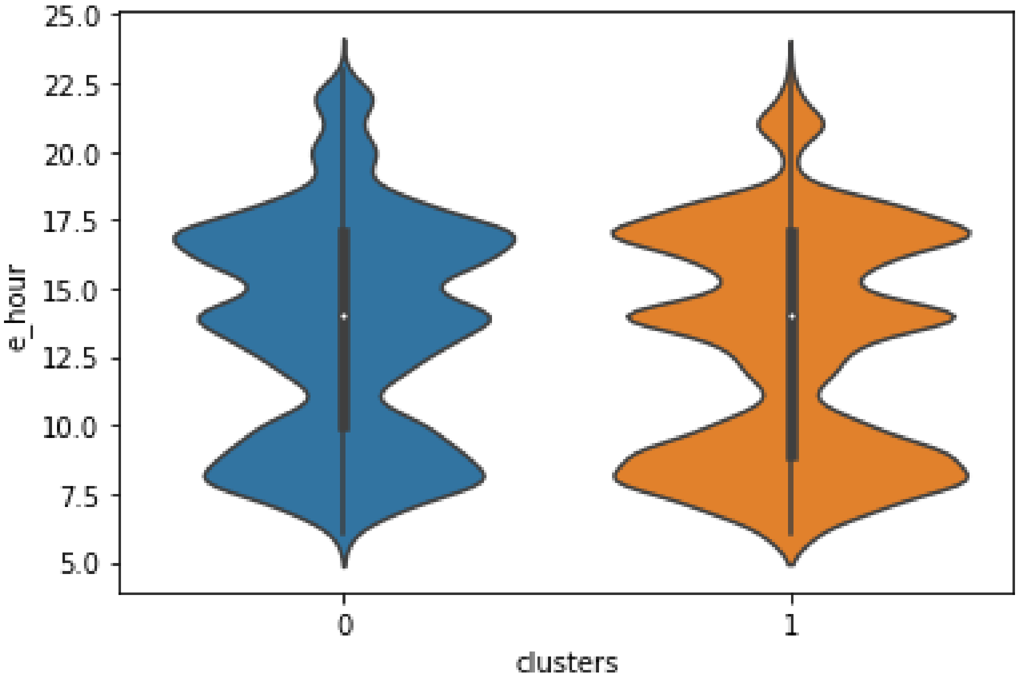


Figure 10: Violin plot of accidents

Accidents Risk Map

Our results as in figure below show that certain tracks become prone to accidents due to factors like over speeding in bad weather conditions and/or at junctions.

Figure 11: Accident Risk Map

Analysis of Lucky Escapes and Overspeeding points

As seen in previous sections of this report, speed plays an important role in traffic safety and therefore makes it vital for this research. To be able to measure the deceleration points, speed value of each point was compared to that of the previous points and only points with negative values were considered as those signified deceleration points. It is normal for a driver to reduce a certain speed while driving but in cases where the drop in speed is sudden and abrupt, we then consider these points as points of interest because the track data collects information from the car every 5 seconds and a sudden reduction in speed of more than 20km/h in 5 seconds can be as abnormal. The values for points of reductions in speed were then grouped into 3, ranging from 0 – 10, 10 – 20 and 20 and above. These points were then visualized on a map in order to have a better understanding and it was discovered that the points in which there was a sudden reduction of speed of more than 20km/h were mostly closed to intersections which are also places which have been confirmed to be accident prone in literatures.

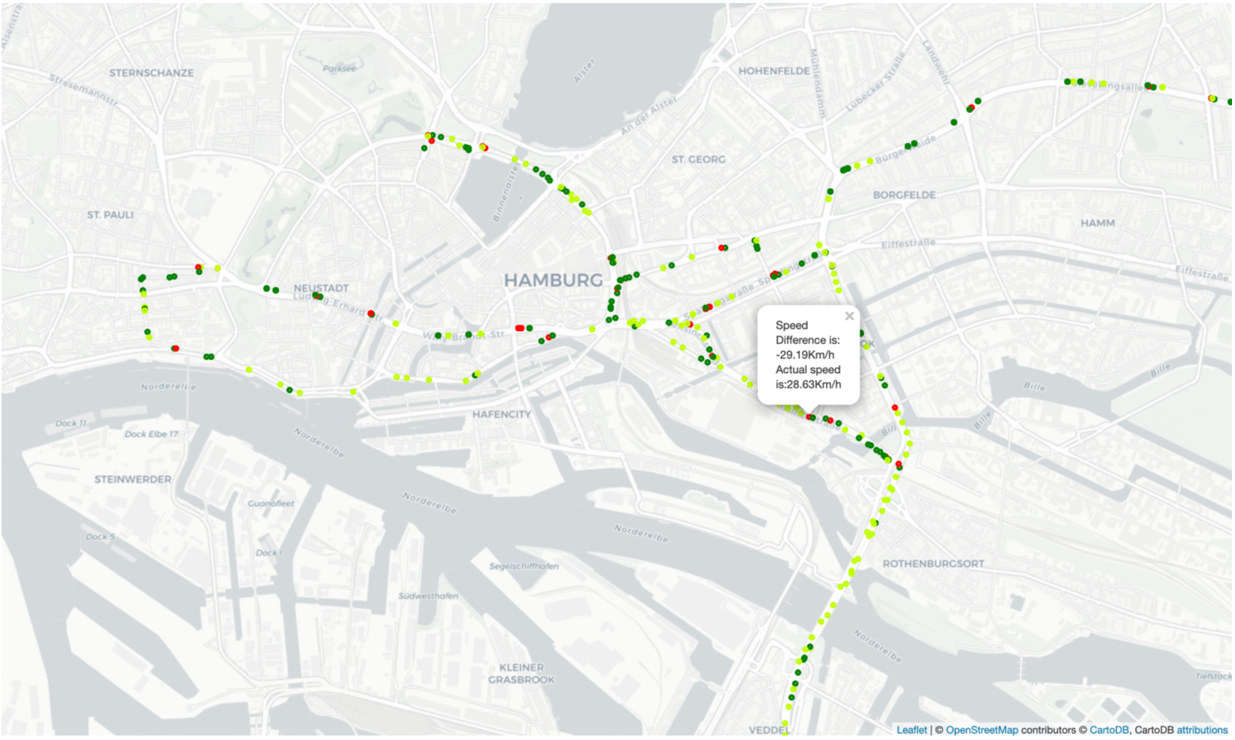


Figure 12: Deceleration Points, the green points are reductions of 0 – 10km/h and the yellow points are between 10 – 20 km/h and the red points are 20 km/h and above

In the figure 12 the red points refer to the points where vehicle speed is above the recommended speed and the yellow ones are the points in which the vehicle speed is within the recommended road speed limit. The accident hot and cold spots, identified from the Getis-Ord method, were discovered to be in spots where vehicle speed was above the recommended speed limit.

The figure 13 shows a map of the study area visualizing the accident hot and cold spots along with the points where vehicle speed exceeded recommended speed limits. The black points represent the accident hot spots, the blue points represent the accident cold spots, the red points represent the point in which the vehicles speed was above the recommended speed limit and the yellow are those within the recommended speed limit.

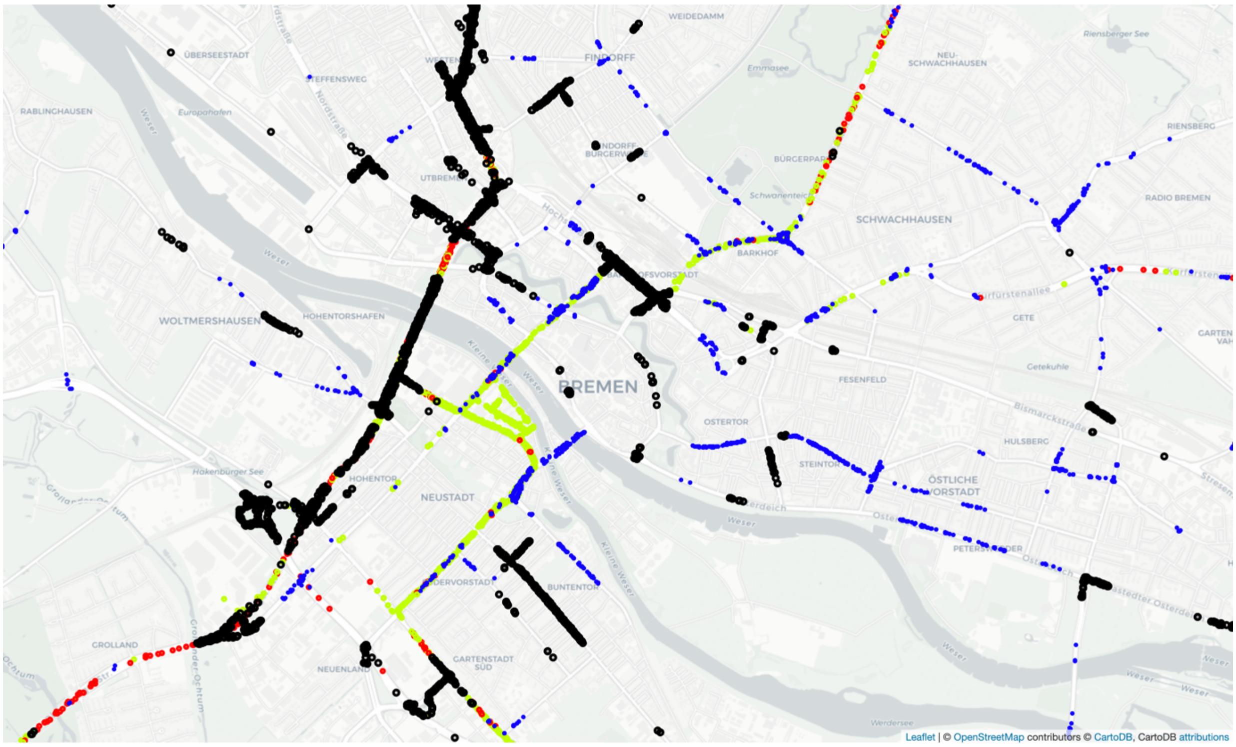


Figure 13: Hot and Cold Spots

Overall from figure 12 and figure 13, we can observe most points of sudden deceleration concentrated around the junctions/intersections. The accident hotspots and cold spots were close to the points of over speeding.

3.6 Modeling and Prediction

Three neural networks were developed to analyse the probability of severe accidents. classification of accidents and future trends of accidents.

Prediction of probability of severe accident

Unfallatlas data of Hamburg was used to develop a binary Artificial Neural Network (ANN) model The model accuracy was found to be 89.41% and confusion matrix to be 89.53% with the loss of 0.34. This shows that the model is in a balanced state between overfitting and underfitting. Similarly other measures like precision, recall, f1 score and kappa were also computed. Mean accuracy of the model was also found near to 90% with variance to be less than 0.1%.

This model was used to test some imaginary scenario under certain criteria. In one case, it predicted that the tendency to die in an accident was zero. When harsh conditions (slippery road, daytime etc) were given it predicted the probability to be nearly 13% compared to 10% of severity when the condition was a dry road in the morning. Though stats showed that the model was performing optimum, we believe that there is always a room to improve the model. This was implemented with a dropout regularization from 0.1 through 0.5. The model was also fine tuned using a grid search method to explore for the better hyperparameters of the model. It was discovered that batch size of 2, epochs of 30 and adam optimizer were the best combination for the designed model.

Classification of accidents

The finalized Euska data for the city of Bremen was used to improvise the previously built network into multi-class model. Before compiling we checked the configuration of the designed model and found that there were a total of 287 trainable parameters in four layers. To quantify the performance of the model, it was evaluated using different statistical measures including mean and variance. The model accuracy was found to be 81.60% and confusion matrix to be 82.00% with the loss of 0.57. This shows that the model is in a balanced state between overfitting and underfitting but there is still some room for improvement. Similarly other measures like precision, recall, f1 score and kappa were also computed and found to be as in the figure below:

Classification Report

	precision	recall	f1-score	support
Class 1	1.00	1.00	1.00	61
Class 2	1.00	1.00	1.00	1315
Class 3	1.00	0.00	0.00	127
Class 4	1.00	0.00	0.00	987
Class 5	1.00	0.00	0.00	142
Class 6	0.77	1.00	0.87	4385
Class 7	1.00	0.00	0.00	38
accuracy			0.82	7055
macro avg	0.97	0.43	0.41	7055
weighted avg	0.86	0.82	0.74	7055

Figure 14: Classification Report

The precision values show that it accurately predicted all classes except class 6.

Prediction of future trends of accidents

The RNN model was used to make predictions for upcoming 275 days (truth data only available for 275 days) and verification done by plotting the line plots with the original number of accidents. The result found is shown in the figure below.

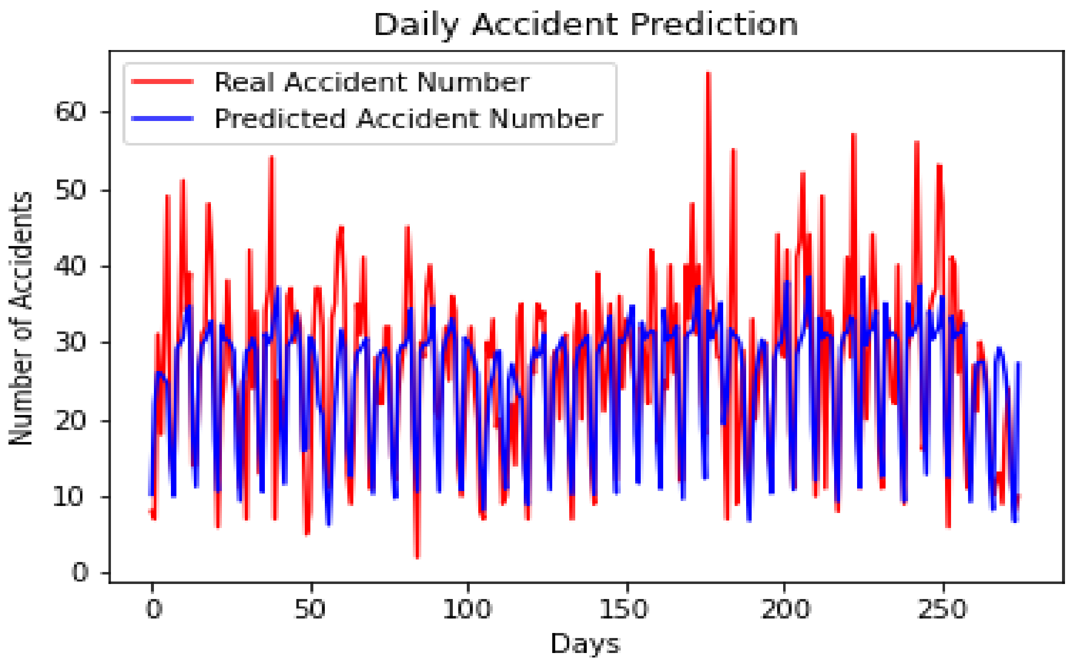


Figure 15: Prediction of daily accidents (for 275 days)

Similarly, the root mean square error was found to be around 9, mean value of daily accidents to be 25 and the relative error below 3.

4 Conclusion

All in all, an exhaustive research of past accidents data and integration of floating car data with open street map and weather data were done to develop insights on traffic safety. Data preparation and engineering were done to convert the data into usable format. We derived a lot of insights on different variables responsible for accidents. These data were then used for exploratory statistical analysis in order to derive the important variables responsible for accidents. Upon processing the data in terms of spatial-temporal analysis and clustering, it was found that junctions were the most vulnerable places of accidents. Overspeed and lucky escapes maps also depicted the abrupt change in the speed at different junctions. However, the reason for this abrupt change may be due to other factors like traffic signals also. Hence a thorough analysis is needed before reaching any conclusion. In addition black spots, hot spots and cold spots were also mapped in relation to pub amenities nearby to see the relation with accidents. Finally, few variables obtained from the exploratory statistical analysis were taken for modeling to compute the probability, categories and trends of future accidents.

References

Bülbül, H. İ., Kaya, T., & Tulgar, Y. (2017). Analysis for status of the road accident occurrence and determination of the risk of accident by machine learning in Istanbul. *Proceedings - 2016 15th IEEE International Conference on Machine Learning and Applications, ICMLA 2016*, 426–430. https://doi.org/10.1109/ICMLA.2016.185

Chen, Q., Meng, Z., Liu, X., Jin, Q., & Su, R. (2018). Decision variants for the automatic determination of optimal feature subset in RF-RFE. *Genes*, *9*(6). https://doi.org/10.3390 /genes9060301

David, E. (20o8). The *Chi Square Statistic* (J. Ryan, Ed.). Mathbeans Project. Retrieved June 12, 2020, from http://math.hws.edu/javamath/ryan/ChiSquare.html

Destatis. (2020, July 14). Driver-related causes of accidents involving personal injury in road traffic. Retrieved July 16, 2020, https://www.destatis.de/EN/Themes/Society-Environment /Traffic-Accidents/_Graphic/_Interactive/traffic-accidents-driver-related-causes.html

Ding, T., Han, X., Han, X., & Zheng, L. (2010). Simulation and analysis of typical winter road traffic accidents in China. *2010 International Conference on Intelligent Computation Technology and Automation*, *ICICTA 2010*, *3*, 315–317. https://doi.org/10.1109/ICICTA. 2010.773

GoodData Documentation. (n.d.). *Normality Testing - Skewness and Kurtosis*. https://help.gooddata.com/doc/en/reporting-and-dashboards/maql-analytical-query-language/maql-expression-reference/aggregation-functions/statistical-functions/predictive-statistical-use-cases/normality-testing-skewness-and-kurtosis#:~:text=As%20a%20general%20rule%20of,the%20distribution%20is%20approximately%20symmetric.

Ishtiaque Ahmed. Road infrastructure and road safety. Transp. Commun. Bull. Asia Pacific, (83):19–25, 2013

Jeon, H., & Oh, S. (2020). Hybrid-recursive feature elimination for efficient feature selection. *Applied Sciences (Switzerland)*, *10*(9), 1–9. https://doi.org/10.3390/app10093211

Thomas, I. (1996). Spatial data aggregation: Exploratory analysis of road accidents. *Accident Analysis and Prevention*, *28*(2), 251–264. https://doi.org/10.1016/0001-4575(95)00067-4

Younus Maqbool. Road safety and Road Accidents : An Insight Road safety and Road Accidents : An Insight. Int. J. Inf. Comput. Sci., 6(4):93–105, 2019.