Urban Road Traffic Accidents:

Comparing risk patterns in Manchester and Birmingham

Abstract

RTA is a major challenge in the world and is increasing at accelerating rate due to modern transport. This is more concerning in urban areas where there is greater population and traffic volume. Mesquitela et al (2022) stated in their study that the accidents are non-random events that happen for reasons at particular locations. However, it is difficult to identify these reasons and predict accidents as there are several factors involved.

WHO (2023) estimated that there are approximately 1.19 million deaths per year globally. This has made traffic accident prediction a popular area in research to prevent further accidents in the future. This dissertation is an attempt to contribute to this field by addressing major gaps in the literature.

The objective is to compare crash risk in Manchester to that in Birmingham during 2019-2023. For this, five major data sources were integrated and used for of spatio-temporal analysis, network modelling, and machine learning methods. The findings found different risk patterns of RTA in each city, highlighting the need for more such comparative studies in the future.

Keywords: RTA, traffic accident prediction, crash risk, spatio-temporal analysis, network modelling, machine learning, risk patterns, comparative studies

Table of Contents

List of Abbreviations

Road Traffic Accidents	RTA

1. Introduction

1.1 Research Subject

This dissertation contributes to the field of traffic accident prediction. The main objective in this field is to identify high-risk locations, also known as 'crash hotspots' (Srikanth and Srikanth, 2020). This is to improve road safety through localised interventions. Various methods have been used for identifying these hotspots and understanding the factors influencing accidents. Some of these popular methods are generalised linear models, time series methods, machine learning, and decision trees (Abdulhafedh, 2017). Neural networks and deep learning have gained popularity in the last years and are currently being explored in the research (Li and Chen, 2025).

One of the significant challenges in predicting RTA is that the degree of impact varies according to location and time. This interaction of spatial and temporal factors cannot be ignored if the objective is to analyse risk or predict the crash hotspots (Alsahfi, 2024). However, traditional studies are limited as they fail to account for the temporal dynamics and spatial heterogeneity of cities (Xia et al, 2024).This issue is concerning in urban cities as the road infrastructure, weather conditions, and several other external factors vary from place to place (Pacheco et al, 2023).

While most of the studies in the field are either country or city specific, this study offers a unique perspective in the field of traffic accident prediction through comparative analysis. In the United Kingdom, Manchester and Birmingham remain at the top in terms of urbanisation, traffic volumes, and population. Both cities are structurally different and have been at risk for traffic accidents for a long time. For comparative modelling, both cities are ideal as the findings can be used to assess the generalisability of predictions and demonstrate the importance of research being location and time specific. In addition, both cities are targets for urban transport policy reforms, making the study more relevant and practical.

1.2 Research Aim and Objectives

This dissertation aims to compare and explore urban road traffic crash risk and predict factors of RTA in two major urban cities in the UK.

For this, four major objectives are addressed:

Objective 1: To analyse the spatio-temporal distribution of reported RTA through charts, KDE, and ST-DBSCAN.

Objective 2: To compare structural differences in both the cities while controlling for traffic volume using network-based modelling.

Objective 3: To identify the key variables influencing RTA in both the cities through multivariate regression, geographically weighted regression, and random forest, and explain the interpretability through SHAP values.

Objective 4: To develop an XGBOOST model and conduct city-specific evaluation and cross-evaluation to test for transferability.

These four objectives allow for a comprehensive analysis and the findings will help to provide policy recommendations that are specific for both the cities. As this is the first such comparative RTA study in the UK, this process also justifies the need for further such analysis.

1.3 Scope and Significance of the Study

This dissertation investigates the crash risk in Manchester to that in Birmingham over a five-year period (2019-2023). The risk can be influenced by many factors, but some key variables were selected. Five major data sources were integrated: STATS19, Annual Average Daily Flow (AADF), OpenStreetMap (by OSMNX), Lower Layer Super Output Area (LSOA), and historical weather. Due to this, the final dataset contains spatial, temporal, geographical, and environmental variables for multivariate analysis. In future, other factors such as vehicle attributes, driver characteristics or taxi trip characteristics can be combined with similar framework.

Understanding patterns of crash risk in urban areas is crucial for improving road safety (Musingura et al, 2023). This is because local agencies and authorities can take informed decisions based on the analysis to implement interventions in locations where accidents have actually occurred (Abdalazeem and Oke, 2025). This dissertation is different from traditional studies that focuses on single geographical area. This is important because comparative analysis shows how diverse road characteristics, traffic volumes, or environmental variables can affect the rate of crashes. Many such possibilities could be explored for analysis of this kind, between two countries or rural cities, for more broader analysis.

1.4 Structure of the Dissertation

This dissertation is structured into four different chapters:

The first chapter is the literature review. It critically examines the existing research in the field of traffic accident prediction. Basic concepts and definitions, the common methodologies that are being used, influencing infrastructural and environmental variables in the field, and the extent of comparative studies conducted are explained in detail. The current gaps in the research field are identified, and a theoretical framework is provided for the analysis. The second chapter deals with the methodological approach. It describes the regions in this study, data sources, and preprocessing techniques. It further explains the methods, statistical models, and machine learning techniques used for the analysis. The third chapter presents the results and the key findings of the analysis. The fourth and final chapter provides policy recommendations, limitations of the study, and proposes directions for future research.

2. Literature Review

2.1 Introduction

Urban transport is being continuously evolved with research still being in progress. Newer options like drones, self-driving cars, and electric vehicles are being explored through technological innovations. While there are clear advantages of such transport, the huge risk of RTA cannot be ignored. This makes assessing such risk crucial to prevent further deaths, injuries, and property damage by developing strategies to improve traffic efficiency (Alpalhão et al, 2025). As discussed, multiple factors are involved in predicting RTA. Hence, these factors need to be efficiently handled for accurate results (Jin et al, 2024). For this, different data sources needs to be integrated to identify such factors and develop targeted interventions. This would help to address the multifaceted problem of crash risk and increase the accuracy of outcomes (He, 2023). The current research in road safety suffers from this limited integration of data sources or combining traditional models with newer machine learning techniques (Skaug et al., 2025). A study very similar to this dissertation is by Vincent et al (2025), where traffic accidents are predicted across six cities, including Manchester and Birmingham using deep learning. The framework is considered to be a novel approach in the field and the results achieved a high accuracy. One of the reasons for this is the integration of different sources so that the dataset used would include different spatial, temporal, and environmental variables.

Most of the studies focuses on a single city for risk analysis. However, these fail to account for local disparities. Comparative studies of the cities makes it possible to identify the differences in organisation and structure that would affect the results of road safety analysis. (International Transport Forum, 2018). Manchester and Birmingham are chosen for this study due to their differences in urban density. This literature review presents the basic theories in the field, the importance of comparative studies and the extent to which it has been conducted, spatio-temporal modeling and machine learning techniques used in the field, and the various infrastructural, environmental, and other influencing external variables and the need to include them. It is concluded by presenting the major gaps in the existing literature and the theoretical framework of traffic accident prediction.

2.2 Theories in the Field

In the literature by Butt and Shafique (2025), the evolution of crash risk has been described in detail. This is from traditional statistical models like Ordered Probit to newer complex models like neural networks that can analyse large datasets and provide real-time feedback. The main emphasis is on integrating data sources to include variables like driver behaviour, weather conditions, and traffic flow to improve the risk prediction. Sun et al (2020) proposed a quantitative crash risk assessment method for the highway authorities to improve road safety. In this paper, the traffic crash risk is defined as the sum of crash frequency and severity. This makes the risk assessment and prediction of RTA to be based on the three: crash frequency, rate, and severity.

Two theories that need to be explained for this dissertation are the exposure theory and the traffic flow theory. Merlin et al (2023) defines exposure as the time a person spends on travel that would increase the crash risk. This exposure would measure the opportunity of an RTA occurring and that crash risk can only be accurately estimated if it is properly controlled. Also, this exposure differs according to the built environment. While this study refers to exposure as an amount, in contrast, Pei et al (2012) interprets it in terms of speed. Speed is one of the contributing factors of RTA. High speed reduces the time of travel and thus reduces exposure. If combined with the study by Merlin et al(2023), high speed decreases the risk of RTA. However, it is a well-known fact that speed is one of the contributing factors of accidents. Hence, the terms in traffic studies differ according to how the researcher interprets, which affects the analysis.

In traffic flow theory, one of the well-known foundational papers is by Greenshields (1934). He proposed that as traffic density increases, the speed decreases. This can be connected with the exposure theory. Traffic density increases exposure and thus would be an influencing factor. The dataset used in this study includes variables from both the theories. It includes the number of vehicles, traffic flow, and the type of road for capturing traffic flow and spatial, temporal, and environmental variables for reflecting factors that would affect the exposure. These can be used for preventing RTA. Haghani et al (2023) proposes a number of tools for a proactive management using crowd risk assessment. While the approach is highly beneficial, the aspect of human behaviour makes the implementation more complex.

In terms of modelling, this study is similar to Crime Pattern Theory. While first developed for analysing crime, this technique is now applied in various fields, including traffic accident prediction. This theory explains that crime tends to get concentrated in places like ‘hotspots’. Hence, this similar idea is used for crash hotspots. This is used in their study by Giménez-Santana et al (2018), for identifying environmental factors that influence RTA in Cadiz, Spain. A crash risk is likely where there are influencing factors and thus this framework is crucial for urban cities like Manchester and Birmingham where they play an important role.

2.3 Cross-City Studies in Road Safety Research

Most studies in road safety research in the UK are conducted with national level datasets. For example, Saini et al (2021) analysed the traffic accidents in the United Kingdom during 2005-2015 to identify patterns and predict crash variables. Though methodologically sound, studies like these are often insufficient as the findings are not generalisable to other regions (Yalamanchili, 2024). Çepni (2017) identified crash risk locations through GIS techniques to establish a relationship between infrastructural variables and traffic accidents. This research demonstrated how urban design, traffic volumes, and environmental variables contribute to the crash risk. While studies like this are helpful for understanding the significance of influencing factors, cross-city studies help with understanding how traffic accidents differ spatially due to such factors. However, the choice of cities is important. Similar road systems but different urban structure and socio-demographic conditions make a comparison of cities more insightful.

Furlong et al (2025) assessed the impact of low-traffic neighbourhoods on RTA in London. The study demonstrated that such a localised approach could significantly reduce road injuries for a safer urban environment. In contrast, Xia et al (2024) provided a more nuanced approach by focusing on spatiotemporal instability. While the former focuses on the effect of urban strategies, the latter identifies specific factors. Both the findings challenge traditional models by supporting that cities have different spatial dependencies and temporal factors. The studies are region-specific and can be applied locally. Hence, they show the value of equity-oriented policies in urban road safety. However, very limited studies have employed direct comparative analysis between two cities.

2.4 External Factors of RTA

Most of the studies have demonstrated that external factors such as weather conditions, infrastructure, or driving speed influence RTA (Tang et al, 2025; Chen et al, 2025). In their review of road crash analysis, Skaug et al (2025) stated that traffic crashes arises from complex interactions between road infrastructure, human behaviour, technological systems, and environmental factors. This makes the exclusion of these a major limitation of several studies. For example, Ren et al (2018) developed a spatio-temporal model to forecast traffic risk but criticised the lack of external variables like weather or road network, which limited the performance of the model. Also, some studies have provided evidence of nonlinear relationships between traffic volume and crash risk (Liang et al, 2024; Carrodano, 2024). This makes the factors much more serious in urban areas due to the high traffic volume and the behaviour of drivers. Also, their distribution is dependent on the built environment (Stiles and Miller, 2024).

Xia et al (2024) developed a spatio-temporal model using a comprehensive real-time dataset across three different urban cities in the United States to study the relationships between crash risk, weather conditions, and geographical attributes. The effects of these external factors on the crash severity of each of the three cities are distinctive. This proved the spatio-temporal instability of the risk factors. It can be concluded through this study that the spatial heterogeneity of urban environments requires specific recommendations rather than uniform policies.

Also, the role of temporal factors is often underestimated in the crash risk studies. To give an example, crash risk during early morning and peak hours is not the same. Gedamu et al (2025) concluded that crashes during nighttime are much more dangerous with the severity differing according to location. Also, crashes more likely to happen on weekends than weekdays. Shikder et al (2025) used time series forecasting to predict RTA during peak hours at urban intersections. The likelihood of accidents is found to be higher during early morning and early evening rush hours. In addition, poor weather conditions such as wet roads or greater wind speed can add to the risk of a crash. This is also confirmed Harith et al (2019) in their research paper that reviewed a total of 45 studies. Adverse weather and nighttime driving were the common factors influencing RTA. Hence, by ignoring all these in the study, analysis would be affected substantially.

2.5 Spatio-Temporal Analysis

Dong et al (2020) explains spatio-temporal analysis as analysing data with respect to its absolute and relative positions in space and time attributes. In simple terms, this technique involves both spatial and temporal variables. Hence, it is highly important and widely used in fields where analysis needs to be location and time sensitive for accurate results. For example, the data used in traffic accident prediction contains spatial and temporal variables and ignoring them would significantly affect the analysis (Koutsaki et al, 2025). Understanding this significance, Choudhary et al (2025) conducted the first spatio-temporal analysis of crash hotspots in India using kernel density estimation. In the case of urban studies, Yuan et al (2012) laid the foundation for spatio-temporal analysis by developing a framework to identify functional regions in a city. His research emphasised the potential of spatio-temporal approaches in urban studies.

While traditional studies focuses on static models and heatmaps, they fail to capture spatial and temporal factors. Hence, this study uses Kernel Density Estimation (KDE), Spatio-Temporal Density-Based Spatial Clustering of Applications with Noise (ST-DBSCAN), and Geographically Weighted Regression (GWR). In a survey of clustering methods, Shi and Pun-Cheng (2019) explain KDE as a cluster with minimum points with ST-DBSCAN extending the method for spatio-temporal data. This dissertation uses a hybrid approach with using KDE to identify high-risk RTA zones and ST-DBSCAN for detecting clusters with data like the location name and the time period of its occurrence.

Spatio-temporal modelling is considered to be powerful in urban traffic studies due to its ability to identify hotspots at specific times and locations (Jin et al, 2024). Hence, these models have played a critical role in explaining local crash risks, including the studies conducted in the United Kingdom. One of the novel approaches in the field is by Chadhuri et al (2023). Using a network triangulation in a spatio-temporal Bayesian model, traffic accident risk was predicted on urban road networks in London. Though methodologically robust, one of the limitations of this study is the exclusion of traffic flow and temporal variables. This has been included in this study.

2.6 Methodologies Used in Traffic Accident Prediction

Several methods have been used in the field of traffic accident prediction. The evolution of these methodologies has been described by Skaug et al (2025) by a systematic review of road traffic accident prediction based on data from multiple sources. The very first technique, or the base model, is statistical learning. Grounded in assumptions, it helps to find patterns in the data and explore relationships between various variables. Popular methods in this include Bayesian models, time series, linear regression, and logistic regression. The common limitations of all these methods are their limited ability to handle high-dimensional data and their sensitivity to outliers. Hence, machine learning provides a clear edge over these methods in its ability to handle huge amounts of data and capture complex, non-linear relationships without any assumptions. In the past research, decision trees, random forests, gradient boosting, support vector machines, and naïve bayes have been used for making predictions. These methods helps to identify high-risk areas and periods and improve road safety (He, 2023). In most studies like Wahab and

Jiang (2019), machine learning models outperform traditional models in predicting crash outcomes.

Neural networks and deep learning are currently being explored and have brought changes in the research field. One of the recent papers that has been highly influential is by Li et al (2025). In this paper, a deep learning model was developed that combines Convolutional Neural Networks (CNN), Long Short-Term Memory networks (LSTM), and Graph Neural Networks (GNN) for predicting RTA in London. They stated that inclusion of spatial and temporal factors increased the accuracy significantly. One of the limitations of this model is the lack of interpretability to understand the predictions. Another study is by Kumar et al (2025), where a hybrid model was developed using random forest, XGBOOST, and Long Short-Term Memory (LSTM) to emphasize integrating multiple machine learning techniques. This is for classification tasks like this study.

This study uses random forest for identifying important features and XGBoost for predictive modelling in both cities. The combination of random forest and XGBOOST is used as a hybrid strategy for a more robust performance and provides both interpretability and accuracy. This strategy balances one of the significant trade-off challenges in the field: the random forest reduces the variance whereas the XGBoost minimises bias. This is critical in urban studies due to the high variability in datasets. In his study, Jamal et al (2021) analysed crash risk factors using various ensemble methods and concluded that random forest performed the best for severe traffic accidents and variable importance. One of the other notable studies is by Chen et al (2025), in which a novel model was developed by combining Modified Stochastic Crested Porcupine Optimiser (MSCPO) with XGBoost to predict traffic accident severity. The model achieved a high accuracy of 83.57% and outperformed other models. Some of the key limitations include the lack of interpretability using SHAP or LIME and training the dataset in different regions for comparison. Both of these limitations have been addressed in this study. Panda et al (2023) discussed the importance of SHAP analysis for predictions, especially for models that are considered black boxes like ensemble methods.

2.7 Theoretical Framework of the Study

This dissertation is based on the exposure theory and the traffic flow theory. The framework is similar to Crime Pattern Theory in spatio-temporal analysis and machine learning prediction. Based on all these, the following question is answered: ‘how do different factors contribute to the crash risk in each city?’.

In cross city studies, normalisation is critical for consistency. This is because characteristics like population or traffic volume differ and making assumptions rather than controlling them would distort the analysis. For exposure, data is normalised through feature engineering. The variables *crashes_per_km, crashes_per_1000_vehicles,* and *crashes_per_10000_vehicles* are created.

The dependent variable in most of the studies is either the frequency or severity of the accidents. The dataset in both the cities of Manchester and Birmingham are highly imbalanced in terms of accident severity. Hence, *accident_count* is the dependent variable in this dissertation. This is ideal since the objective is to predict the crash risk. It allows for both exploratory analysis and predictive modelling.

A good theoretical framework for modern traffic safety integrates multiple factors. The equation used in this dissertation is:

Crash Risk (R) = f(Spatial and Exposure Factors, Temporal Factors, Infrastructural Factors, Environmental Factors).

Below are the major variables used in this study:

Spatial and exposure factors:

SOA21CD, LSOA21NM, LSOA21NMW, BNG_E, BNG_N, LAT, LONG, geometry, centr oid, city,avg_daily_traffic, avg_daily_traffic_idw, has_traffic, accident_count, avg_severi ty, crashes_per_km.

Temporal factors:

datetime, day, month, hour, accident_year, is_day, date, day_of_week, time.

Infrastructural Factors:

road_length, road_km.

Environmental factors:

temperature_2m, precipitation, snowfall, wind_speed_10m, cloud_cover_low, relative_ humidity_2m.

The above factors are the independent variables used in this study. The selection of these variables depends on each of the objective.

One of the limitations of this study is the exclusion of behavioural factors from the framework. However, research in human behaviour has revealed complex relationships with psychological and demographic factors that challenge safety approaches (Skaug et al, 2025). For example, in the study by Lacherre et al (2024), driving behaviour is emphasized to be a major contributor to traffic accidents. Similarly, Mohammed (2025) analysed 45 behavioural and demographic variables and identified several factors influencing crash risk such as speeding, emotional instability, age and gender of the driver. Hence, for future research, this is to be included in the analysis rather than relying on assumptions.

The main models used in this study are kernel density estimation, Spatio-Temporal Density-Based Spatial Clustering of Applications with Noise (ST-DBSCAN), multivariate regression, Geographically Weighted Regression(GWR), random forest, and XGBOOST.

To summarise, this dissertation combines spatial, temporal, exposure, and environmental factors for spatio-temporal modelling and predictive analysis within the context of Manchester and Birmingham. Though individual insights are driven, the major focus is on comparative analysis to strengthen the argument for the need of such studies for generalisability of findings.

2.8 Existing Gaps in the Field

Although numerous studies have been conducted for contributing to the traffic accident prediction, there still exist some gaps. This dissertation fulfils four critical gaps in the existing literature:

Very limited studies focus on comparative analysis between two cities. This study provides a comprehensive and predictive analysis of the risk of traffic crashes in both the cities between 2019 and 2023. A similar study is by Chen et al (2024) where comparative analysis is conducted in which New York and Chicago were selected to predict risk through a novel multi-granularity hierarchical spatio-temporal network.

Combining multi-source data is critical in traffic accident prediction. However, several studies rely on single or limited datasets. This study integrates traffic accident data with traffic volume, weather, infrastructure, and geographic data to analyse crash risk factors.

Traditional studies often rely on static spatial methods or simple regression models. This study uses spatio-temporal hotspot detection using ST-DBSCAN and KDE for visual analysis.

While machine learning is increasingly used in transport modelling, interpretability often is not explained. This is addressed by applying Random Forest and XGBoost for prediction and using SHAP values to maintain transparency. Integrating SHAP analysis in the evaluation of such models improves the interpretability of meteorological and temporal features (Durap, 2025).

2.9 Summary

A critical literature review has been presented in this manner, focused on the comparative analysis, spatial-temporal modelling, machine learning, and environmental and infrastructural variables in the domain of urban traffic crash risk.

To summarise, individual datasets provide unique insight into the urban traffic crash risk; however, an integration of these sources is required for accurate results and forming effective policy recommendations. There are no holistic, city-comparative models available that combine these sources for specific strategies. The study also fulfils an urgent need of geographically comparative and methodologically hybrid research by determining risk patterns in Manchester and Birmingham over 2019-2023. It forms a theoretical framework on the integration of data sources based on exposure theory, traffic flow theory, and crime pattern theory. In this way, the study adds to the emerging literature that does not just predict crashes but investigates and tries to prevent them through special urban interventions.

3. Methodology

3.1 Overview of Manchester and Birmingham

Manchester and Birmingham are popularly known as UK's twin cities. This is due to its similar characteristics and have been competing as major urban centre after London. Due to this, both are often used in comparative research in urban studies. For this dissertation, both are selected for methodological and practical considerations.

In terms of urbanisation and population density, both the cities remain at the top (City Population, 2025). In addition, both have comprehensive spatial and temporal data in all the data sources. This ensures a complete and balanced framework for meaningful analysis. Despite the similarity, the structural differences of both the cities provides an opportunity to conduct valid analytical research.

From a practical perspective, both the cities are the targets of the Vision Zero Strategy. The aim of this strategy is to deliver a safe system approach for road safety. The objective in Greater Manchester is to eliminate road deaths and life-changing injuries by 2040 (Greater Manchester Combined Authority, 2023). Whereas the Road Harm Reduction strategy in Birmingham has adopted a healthy streets approach for proactive measures towards road harm risk (Birmingham City Council, 2023). Therefore, this study would help these action plans to develop relevant strategies for both the cities. The study will also help to address the need for similar major transport reforms in other urban areas.

3.2 Data Sources

A multivariate spatial model often outperforms univariate spatial model, especially in the field of predicting complex phenomena like road traffic accidents, due to the influence of several correlated factors (Huang et al, 2017) One of the major advantages of this study is the use of integrated data from multiple sources. The following five sources were integrated for a dataset that is rich with spatial, temporal, and external variables for analysis:

STATS19 accident data: This is the official dataset containing collisions reported by the police It basically contains various information on each accident. It is widely used in transport research by several studies (Doudaran, 2020; Lovelace et al, 2020). One of the major limitations is underreporting, that is, the data source only contains the accidents reported by the police.

Average Annual Daily Flow (AADF): This official dataset is published by the UK Department for Transport (DfT) and contains one of the significant variables of traffic accident prediction, traffic flow estimates (Department for Transport, 2024). It is widely used in traffic studies and is reliable. However, it is an estimate and not a direct count. Chang et al (2023) developed machine learning models using this dataset that achieved a high accuracy in risk prediction.

Lower Super Output Areas (LSOA) Boundaries: LSOA is a small geographical area used for statistical analysis. This dataset is provided by the Office for National Statistics website and enables linking the accidents to demographic and environmental features. Similar to this study, Lovelace et al (2020) combined stats19 dataset with LSOA boundaries to analyse risk patterns. A limitation is that the crashes may not align accurately within the boundaries and that there is a risk of reduced risk when fewer crashes occur in a particular LSOA.

Open Meteo weather data: This is an open-source API providing historical weather variables at hourly resolution. One of the significant limitations of this study is that the weather variables in this data source may not be accurate for exact crash locations. However, this is chosen over the official MIDAS dataset because the latter has huge amounts of missing data and a small number of weather stations in the city for accurate analysis. Since the analysis is of two particular cities and not all over the UK, the use of Open Meteo is validated.

OSMNX road network data: This data is obtained from OpenStreetMap (OSM) and is widely used in transport research. In the foundational paper by Boeing (2017), OSMNX python library was introduced for street network analysis in urban planning and transport research. It provides a detailed road network structure and allows for integration with other data sources. While the dataset was checked for completeness, the accuracy may vary as it is a crowdsourced dataset.

Each data source has its own importance and limitations. It is up to the researcher to use the sources carefully while handling the limitations efficiently. In the field of traffic accident prediction, government and open data sources are considered to be crucial. They form a reliable foundation for analysis. This is described by He (2023). According to him, government datasets provide detailed and complete records and open datasets expand the scope of government datasets. This is by providing weather, demographic, and road network data. Hence, these are used for this study after careful consideration.

3.3 Data Preprocessing and Integration

Data preprocessing is crucial for enhancing the quality and usability of traffic accident data. Differences in data collection and management requires rigorous preprocessing to avoid inconsistent results. This includes steps like data cleaning, feature engineering, standardisation, normalisation, or data balancing (He, 2023).

The following preprocessing and integration steps were conducted for the final dataset that has been used in the analysis:

The STATS19 accidents data was loaded and was filtered for Manchester and Birmingham using their ONS codes ["E08000003", "E08000025"]. The date and time columns were combined into a new column called 'datetime'. There were no rows that failed this parsing. The columns that are not needed for analysis were dropped. These include: accident_reference, accident_number, second_road_number, did_police_officer_attend_scene_of_accident, enhanced_local_authority, and enhanced_severity_collision. These columns are not useful both for exploratory and predictive modelling and hence have been excluded from the dataset.

Then, the hourly weather data of both cities were combined into a dataset called 'weather_data' and merged with the accident data as 'merged_data' by aligning by the nearest hour.

The accident data is converted to a geodataframe, and the LSOA data is loaded and filtered for both cities. The CRS of accident data is reprojected to the CRS of LSOA data and spatially joined. Accidents that were not matched were dropped from the analysis. This is because there were only 36 unmatched accidents(0.25%). The accident data is then aggregated with the accident count and average severity. Missing values were filled with 0, which again was only 2.1% of the data.

The road networks of both cities were loaded from OSMNX and converted to geodataframes. Both datasets were merged into a full dataset. The geodataframe was split by LSOA boundaries. The road length was computed and aggregated by LSOA and is then merged with the LSOA dataset. It was checked if there was any LSOA with zero road length, and there were no such LSOAs. The *crashes_per_km* risk metric was computed.

Finally, the AADF data was loaded and filtered for the years 2019-2023. It was then converted to a geodataframe and filtered for both cities. The AADF points were spatially joined to nearby roads within 30 m and then to LSOA polygons. The average daily traffic was computed by grouping LSOAs. This dataset was named ‘aadf_per_lsoa’.It is then merged with the road network data. This is the final dataset and was named ‘final_data’. Exploratory and predictive analysis is conducted with this final data.

LSOAs with missing average daily traffic were imputed by the inverse distance weighting method and were assigned to the final dataset. This method is often supported for sparse datasets in the research field. This is one of the challenges of urban traffic studies, that traffic data is often limited and there is data lost when aggregated for analysis. Though this imputation is a limitation for this study, the traffic volume is only used for exploratory analysis. For predictive modelling, since XGBOOST is used, it handles imputed values better than other linear or traditional statistical models.

4. Results

4.1 Objective 1 - Analyse the spatio-temporal distribution of reported RTA

4.1.1 Kernel Density Estimation for Spatial Distribution

Kernel Density Estimation (KDE) is a non-parametric statistical method. In the field of traffic accident prediction, it is used for identifying the traffic accident hotspots by visualising the probability density estimation of the areas. For example, Srikanth and Srikanth (2020) applied KDE to identify and rank accident hotspots in Des Moines City from 2008 to 2012. Similar to this study, 5 years of accident data were used, and the methodology could be applied in any other city with sufficient data availability.

After filtering data separately for Manchester and Birmingham, KDE was plotted for both the cities.

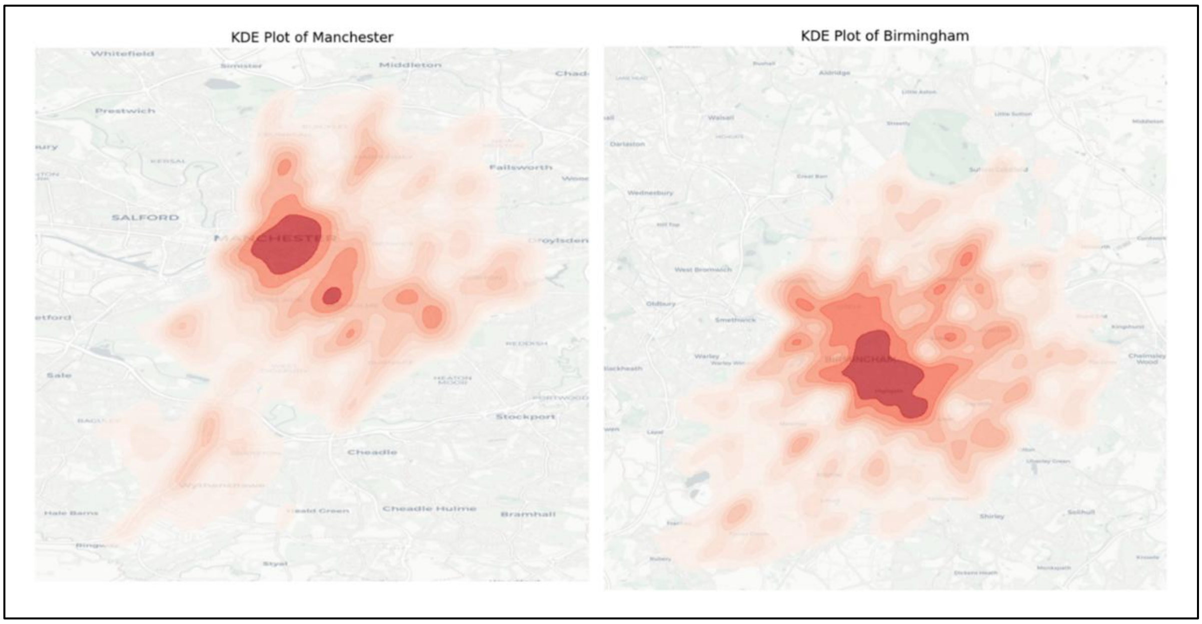

Figure 1: KDE Visualisation of Manchester and Birmingham

Figure 1 shows the KDE plots of the accident data of both cities during the period 2019 to 2023.The darker red areas depict the highest probability density of the accident hotspots in the city. Both cities illustrate the typical core-periphery model of urban development: the central urban core has more development, whereas the periphery areas lack investment. This aligns with the traffic accident risk. Both cities have a high risk in the Central Business District (CBD) – Manchester city centre and Birmingham city centre. Both these areas have high population, business activity, traffic volume, and dense road networks.

Manchester has intense clustering in specific areas. There is less concentration away from the centre and the hotspots are spread in multiple directions. The accidents are distributed to Salford, Hulme, and other southern central corridors. There are small clusters over major transport corridors – commuter roadways, industrial, and university-linked zones (Oxford/Wilmslow Road, Stockport Corridor, Salford, and Ashton Old Road). Peripheral areas like Prestwich, Bramhall, and Chadderton have a low density of hotspots, with high density possibly at intersections.

Birmingham has a more centralised and compact core than Manchester, with less intense and dispersed hotspots across the city. Birmingham's urban structure has the city's major roads radiate from the centre, and the arterial corridors are likely dominated with accident hotspots – A34 (Stratford Road), A38 (Bristol/Kingsbury Road), A45 (Coventry Road), and A4540 (Ring Road).

To summarise, Manchester has a radial pattern that requires targeted policies whereas Birmingham has a concentrated and networked pattern that requires city-wide and systematic strategies.

4.1.2 Charts for Temporal Analysis

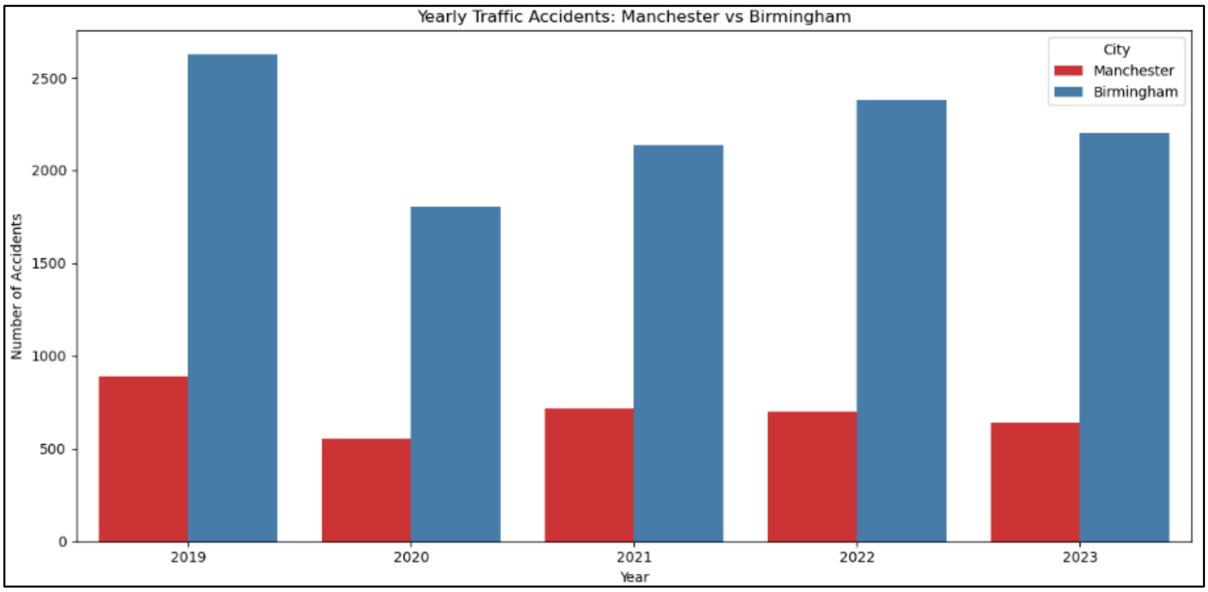

Figure 2: Yearly Traffic Accidents in Manchester and Birmingham

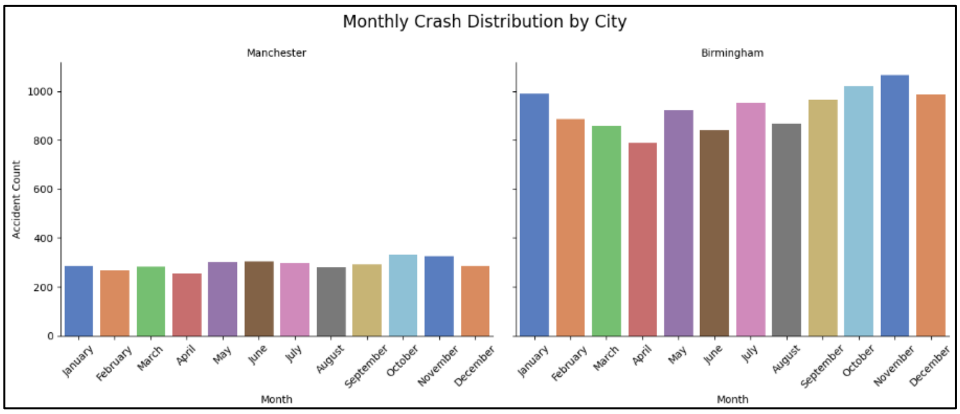

Figure 3: Monthly Traffic Accidents in Manchester and Birmingham

Figures 2 and 3 show the yearly and monthly traffic accidents in Manchester and Birmingham. Both cities show a noticeable decrease in accidents in the year 2020, which corresponds to the COVID-19 pandemic. Following the restrictions on human mobility, both show an increase in the following year. Manchester shows a stagnating or declining trend, whereas Birmingham shows a rising trend. Both cities show a slight decrease in 2023, which could be due to stabilisation, safety measures, hybrid working, or even underreporting of accidents to the police. Overall, it must be noted that Birmingham has a larger number of accidents than Manchester every year. In the monthly bar charts, Manchester illustrates a less minimal seasonal pattern than Birmingham. Birmingham has an increase of accidents in late autumn and early winter.

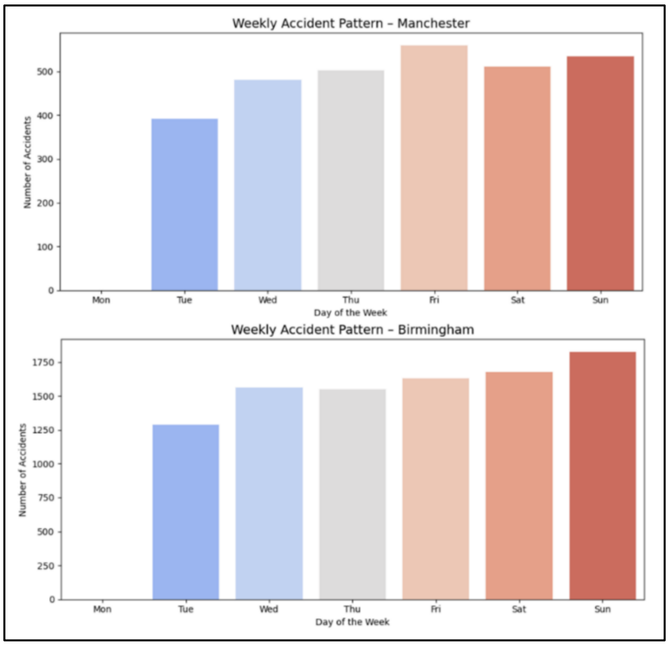

Figure 4: Weekly Traffic Accidents in Manchester and Birmingham

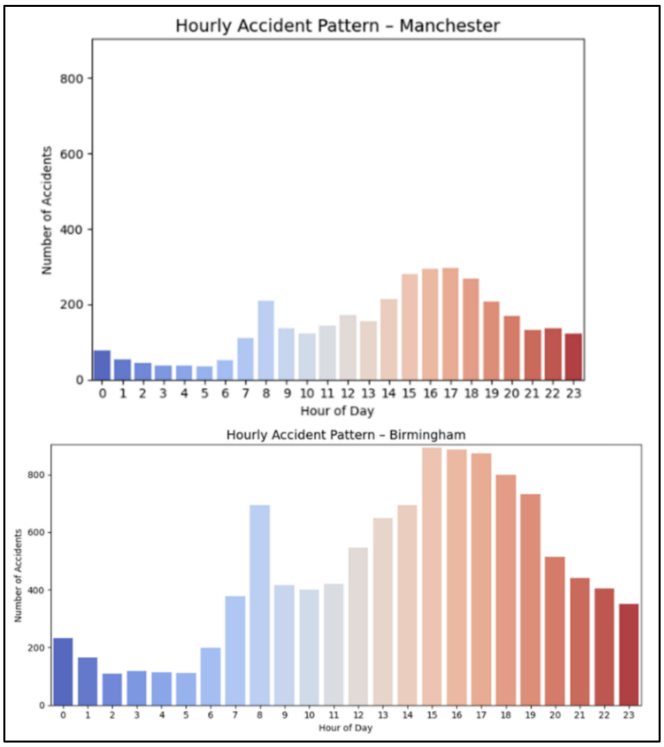

Figure 5: Hourly Traffic Accidents in Manchester and Birmingham

Figures 4 and 5 show the weekly and hourly traffic accidents in Manchester and Birmingham. Both cities show the expected temporal patterns. There is a morning peak at 8.00 am and an evening peak due to rush hour at 15.00-17.00. There are fewer accidents at nighttime due to lower traffic volume. This strongly aligns with the urban commuting behaviour. Manchester follows a typical weekday crash pattern that peaks on Friday. Birmingham shows more weekend risk, with Sunday as the peak. This suggests more weekday rush hour interventions in both cities, particularly in the afternoon. However, the weekend strategies should align with the overall interpretation.

4.1.3 ST-DBSCAN for Spatio-Temporal Hotspot Detection

Birant and Kut (2007) presented a foundational paper in the niche area of spatio-temporal clustering, Cluster analysis is an unsupervised learning method that is popular in the fields of science and engineering like transport, health, or environmental studies for grouping data based on density patterns. ST-DBSCAN is such clustering method that groups based on both spatial and temporal dimensions (Shi and Pun-Cheng, 2019).

Persistent spatio-temporal hotspots are detected in Manchester and Birmingham using ST-DBSCAN. After several trial and error, parameters chosen were: eps1=200; eps2=60;

min_samples=6. This is appropriate for a balanced, realistic urban setting while not being too restrictive.

A cluster summary was created that contains number of accidents, first date of accident, last date of accident, and the list of years where accidents have occurred. This is to give a preview of the cluster to analyse each hotspot. For more information, hotspot locations were then geocoded using Nominatim to get approximate road names. This helps in identifying persistent clusters that have existed for multiple years and hence are at high risk.

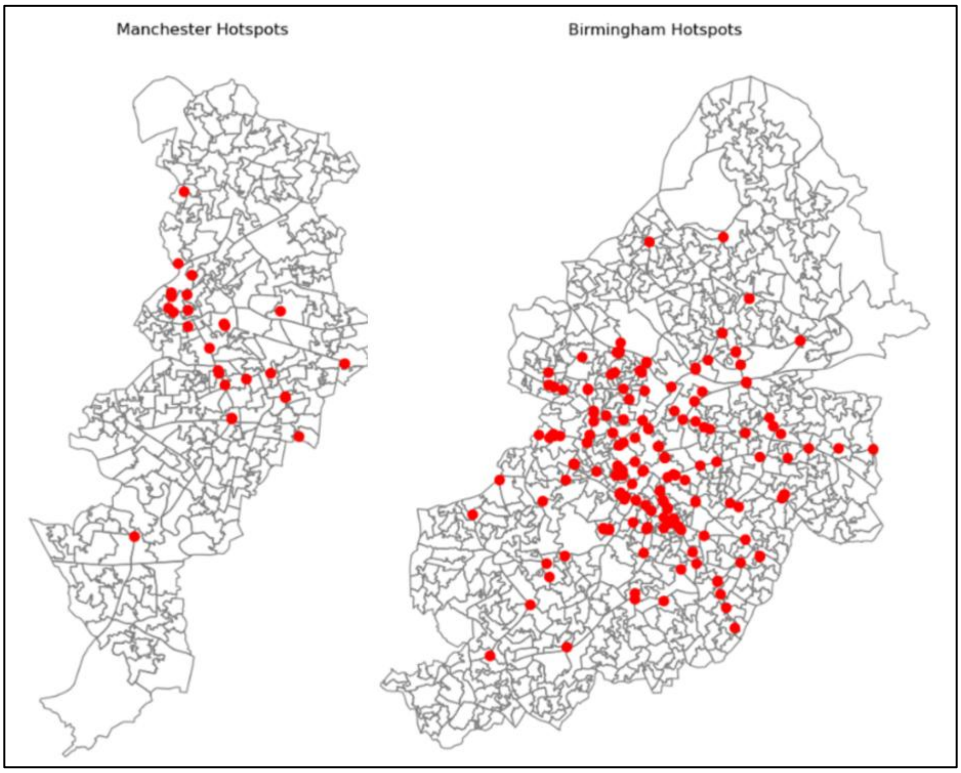

Figure 6: ST-DBSCAN for spatio-temporal hotspot detection in Manchester and Birmingham

The ST-DBSCAN shows a clear advantage over KDE, as hotspots are clustered based on spatial density rather than geographical boundaries, making it more suitable for inter-city analysis. The maps suggest that Birmingham faces a more widespread issue ,whereas Manchester is more spatially concentrated. Deeper analysis for persistent clusters gave more insightful results.

In Manchester, clusters are concentrated at the central and north-central parts in specific neighbourhoods. Wilmslow Road is at the highest risk as the hotspot has been persistent hear for three years. Oxford Road, Stockport Road, and Bradshaw Street has been at risk for past two years. Recent hotspot in 2023 has appeared in Deansgate. Hence, major public transport and arterial roads remain at risk. Hotspots at Trinity Way, Beetham Tower, and Woodlands Road has disappeared and has been absent for three years. Hence, these areas could be further analysed to identify reasons that has reduced crash risk and if it could be implemented in other areas, especially Wilmslow Road.

Hotspots in Birmingham span across most of the city. More persistent hotspots were found here for five years. These include Bordsley Park Road and Haden Circus in the Birmingham Ring Road. Bristol Street, Stratford Road and Nechells are dangerous since hotspots have been present here for three years. Bordsley Street, Garrison Street, Golden Hillock Road, Selly Oak, and Chester Road could be further studied as hotspots have been absent here for two years. This information gives further insights. Roads in suburban and residential areas have become safer whereas roads in major transport areas, commercial and industrial zones remain at risk.

4.1.4. Summary

Metric	Manchester	Birmingham
Total crashes	3498	11144
Crashes per km	1.313284	2.220908
Stable spatio-temporal clusters(2021-2023)	1	14
Highest crash month	October	November

Table 1: Summary Table for Spatio-Temporal Analysis

Birmingham has a substantially higher total number of accidents and crash intensity. It also exhibited far more persistent spatio-temporal hotspots, indicating more prevent high-risk areas. Additionally, the peak crash month differed, with Manchester’s highest in October and Birmingham’s in November, suggesting seasonal variation may be more pronounced in Birmingham.

4.2 Objective 2 - Compare structural differences in both the cities

The study by Neelima et al., 2025 explored how road networks interact with the urban environments and emphasized the need for network modelling through comparative studies.

4.2.1 Structural Metrics Analysis

City	Mean	Median	Standard Deviation
Manchester	1.3566	1.0607	1.143
Birmingham	2.2779	1.6993	1.8679

Table 2: Crashes per km for Manchester and Birmingham

City	Mean	Median	Standard Deviation
Manchester	33.1672	8.374	138.8297
Birmingham	38.3932	14.2556	112.133

Table 3: Crashes per 10,000 vehicles for Manchester and Birmingham

By normalising *crashes_per_km* and per *crashes_per_10000_vehicles*, differences in road network length and traffic volume are controlled, allowing a structural comparison that is independent of exposure.

Birmingham has significantly higher crash density than Manchester, with great variation due to spatial heterogeneity. Both the *crashes_per_km* and *crashes_per_10000_vehicles* show skewed data, suggesting that the crashes are not evenly distributed but concentrated on specific high-risk locations. There is also evidence of outliers and accidents in intersections, junctions, or corridors. Even after adjusting for network length and traffic volume, Birmingham's road segments are riskier than Manchester's, indicating that the risk is more spatially variable due to road network characteristics or urban infrastructure.

4.2.2 Network Structure Comparison

City	Average Node Degree	Average Betweenness Centrality	Average Edge Length
Manchester	4.7105	0.0059	0.0009
Birmingham	4.4292	0.0043	0.001

Table 4: Network Risk Metrics of Manchester and Birmingham

The average node degree shows that Manchester has a denser and more interconnected road network. Also, its lower average betweenness centrality indicates traffic concentration at key nodes. Both have similar average edge lengths, with Birmingham having slightly longer road segments between intersections.

The results of centrality align with the pattern found in KDE, with hotspots in Manchester being more concentrated while those in Birmingham are more distributed.

4.2.3 Spatial Visual Analysis

The spatial analysis of structural metrics revealed distinct patterns in both the cities. The visualisations depict the structural and exposure risk.

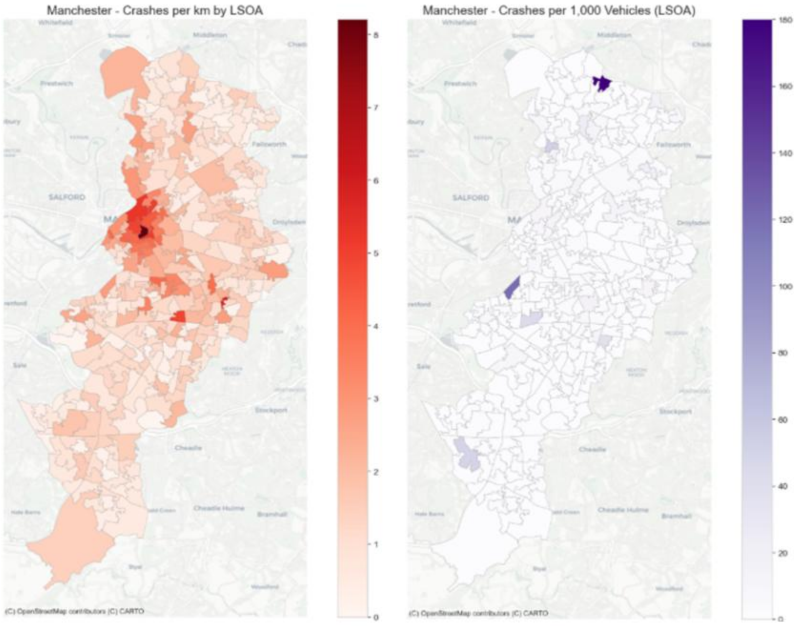

Figure 7: Spatial Visual Analysis of Manchester

Manchester shows concentrated risk on transport corridors, likely the major arterial roads. Also, there are fewer areas affected due to the traffic volume. However, this may not be true risk but rather due to the denominator effect.

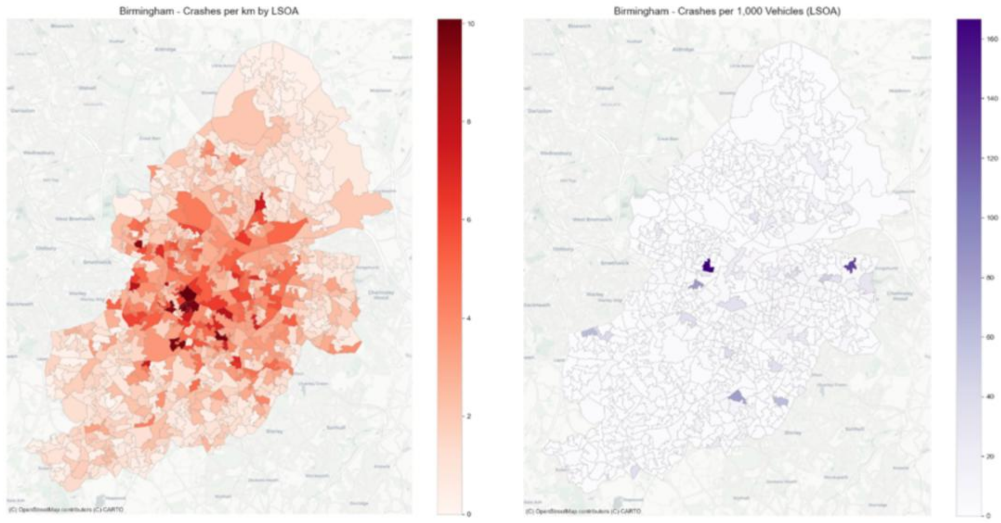
Birmingham - Crashes per km by LSOA
0
2
4
6
8
10
(C) OpenStreetMap contributors (C) CARTO
Birmingham - Crashes per 1,000 Vehicles (LSOA)
0
20
40
60
80
100
120
140
160
(C) OpenStreetMap contributors (C) CARTO

Figure 8: Spatial Visual Analysis of Birmingham

Birmingham depicts the radial risk pattern, and there are several multiple hotspots influenced by traffic volume. This may be driven by larger population density and data availability.

Manchester shows a more linear pattern focused in corridors, while Birmingham covers a large geographic area. The persistent hotspots in Birmingham despite normalising the traffic volume support that the crash risk is influenced by infrastructural and behavioural factors. These spatial patterns support the hypothesis that structural differences in road network design are connected with traffic crash risk.

4.2.4 Welch T-test

Metric	t-Statistic	p-Value
Crashes per km	-9.344	0.0000
Crashes per 1,000 vehicles	-0.587	0.5577

Table 5: Welch T-test of network risk metrics

The t-test gives a strong difference of *crashes_per_km* in both the cities, with Birmingham having higher crash density. By contrast, there is no such evidence for crashes per 1000 vehicles. Traffic volume is a key variable in the crash risk difference of both the cities.

4.2.5 Correlation Analysis

Correlation analysis revealed statistically significant negative relationships between road length and crash density (Pearson r = -0.108, p = 0.001; Spearman r = -0.167, p < 0.001). Shorter segments are generally more crash-prone per kilometre. Similarly, traffic volume exhibited a weak-to-moderate negative correlation with crash risk per vehicle (Pearson r = -0.198, p < 0.001) and crash risk per km (Spearman r = -0.574, p < 0.001), suggesting that higher-volume roads may be structurally safer per unit exposure. These findings strengthen the hypothesis that structural factors such as segment length and traffic capacity influence the spatial crash risk, supporting the need to adjust for them in comparative analysis.

4.3 Objective 3 - Identify the key variables influencing RTA

Several studies use multivariate regression and random forest for risk modelling. However, Geographically Weighted Regression(GWR) is not integrated in the same framework, which is significant for understanding spatial heterogeneity. GWR is particularly important when weather data is integrated to account for spatial autocorrelation. This is the first study that has combined all the three models in the same framework. This helps in understanding why, where, and when RTA occur.

4.3.1 Multivariate Regression

Statistic	Value
R-squared	0.230
Adjusted R-squared	0.223
F-statistic	34.88
P value	0.000

Table 6: Model Summary Table of Multivariate Regression

Variable	Coef(β)	SE	t	p	LLCI (b)	ULCI (b)
const	-43.928	10.088	-4.36	0.000	-63.726	-24.131
road_length	0.0018	0.0001	14.72	0.000	0.0015	0.0020
avg_daily_traffic_idw	0.0001	0.00003	3.87	0.000	0.00007	0.0002
temperature_2m	0.531	0.244	2.18	0.030	0.052	1.009

precipitation	-5.824	2.766	-2.11	0.036	-11.253	-0.396
snowfall	10.257	26.677	0.38	0.701	-42.097	62.610
wind_speed_10m	0.487	0.153	3.19	0.001	0.187	0.787
cloud_cover_low	-0.031	0.032	-0.98	0.325	-0.093	0.031
relative_humidity_2m	0.411	0.101	4.06	0.000	0.212	0.609

Table 7: Multivariate Regression Coefficients

In the case of urban traffic studies, acceptable R-squared values is between 0.50 to 0.99. The lower limit is 0.50 because urban traffic studies includes urban behaviour and external variables that are complex in prediction.

A multivariate regression model was developed with *accident_count* as the dependent variable. While the model is statistically significant (F-statistic = 34.88), it is only able to represent 23% of the variation (R-squared = 0.230).

Road length, traffic volume, temperature, wind speed, and humidity were significant positive predictors; precipitation was a significant negative predictor, while snowfall and cloud cover were not significant.

All variables have a Variance Inflation Factor (VIF) < 2. There is very low multicollinearity among predictors and thus the model estimations are stable and reliable.

4.3.2 Geographically Weighted Regression

While multivariate regression gives an overview, it was not a good methodology for developing models for each city for comparison as the variation it was able to explain was very low. So, the geographically weighted regression model was then developed for each city for comparison.

Variable	**Mean**	**Std**	**Min**	**Median**	**Max**
Intercept (X0)	12.885	3.518	7.073	12.535	20.136
road_length	6.047	1.098	3.808	5.740	8.177
avg_daily_traffic_idw	0.126	0.641	-1.488	0.131	2.109
temperature_2m	-0.181	0.937	-3.212	-0.204	1.793

precipitation	0.316	0.914	-1.030	0.104	2.534
snowfall	2.139	4.287	-1.016	0.038	12.728
wind_speed_10m	0.485	0.665	-0.987	0.400	2.061
cloud_cover_low	0.150	0.704	-1.357	0.137	2.549
relative_humidity_2m	0.361	1.048	-2.277	0.473	2.972

Table 8: GWR Local Coefficients of Manchester

The GWR model for Manchester was able to explain 56% of the variation (R-squared = 0.563). Structural factors have the most influence on Manchester, with road_length being the significant variable. Weather has less and inconsistent effect on the city. Snowfall is highly variable, suggesting some regions are sensitive.

Variable	Mean	Std	Min	Median	Max
Intercept (X0)	19.463	6.588	9.571	18.090	33.706
road_length	9.820	4.617	1.581	9.877	18.004
avg_daily_traffic_idw	3.198	1.985	-1.296	2.999	7.532
temperature_2m	1.008	1.245	-1.794	0.786	5.077
precipitation	-0.159	1.327	-2.893	-0.205	3.746
snowfall	0.501	1.156	-1.659	0.280	3.391
wind_speed_10m	0.357	1.032	-2.292	0.431	2.300
cloud_cover_low	-0.409	0.639	-2.176	-0.390	0.889
relative_humidity_2m	1.662	1.419	-1.127	1.409	5.733

Table 9: GWR Local Coefficients of Birmingham

The GWR model for Birmingham was a stronger fit and was able to explain 69% of the variation (R-squared = 0.69). Traffic volume is the strongest predictor, with humidity being consistently positive. Road length is more variable than in Manchester. Other weather factors are locally significant in some regions.

In summary, Manchester is driven by road network characteristics. Birmingham is more spatially heterogeneous and needs locational strategies. Traffic volume influences the accidents in both the cities. Microclimate effects are seen more in Birmingham than in Manchester, and hence it needs for weather-responsive measures.

4.3.3 Random Forest

Random Forest was developed for both cities with rigorous hyperparameter tuning and log transformation of input features to find the important variables.

Metric	Value
R-Squared	0.4822
MAE	5.21
RMSE	7.62

Table 10: Model Summary Table of Random Forest for Manchester

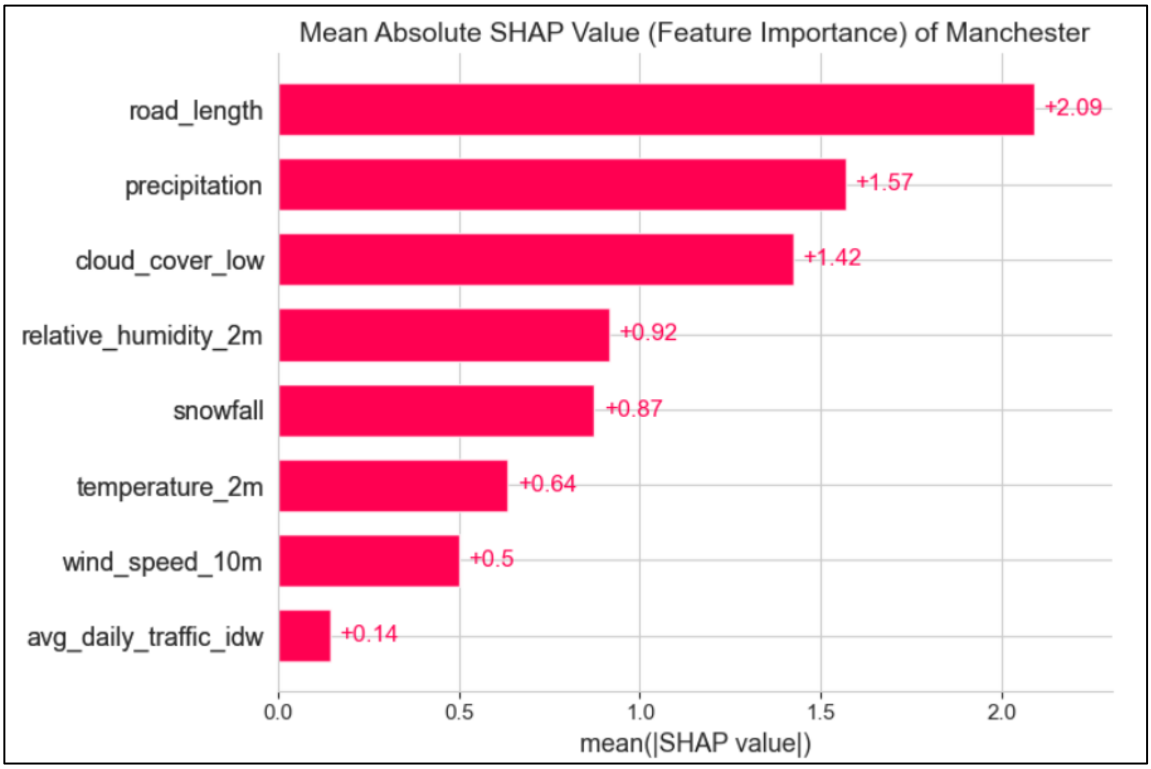

Figure 9: SHAP Value Feature Importance of Manchester

The model shows moderate explanatory power (R-squared = 0.4822). The SHAP values interpretation chart shows important insights into the features influencing Manchester. Road length is the most influential factor, as seen in the GWR model. Precipitation and cloud cover indicate that wet conditions and low visibility influence crash risk. Humidity and snowfall have a moderate impact. However, traffic volume shows much less impact, possibly due to data limitations or variability.

Metric	Value
R-Squared	0.550
MAE	8.19
RMSE	13.19

Table 11: Model Summary Table of Random Forest for Birmingham

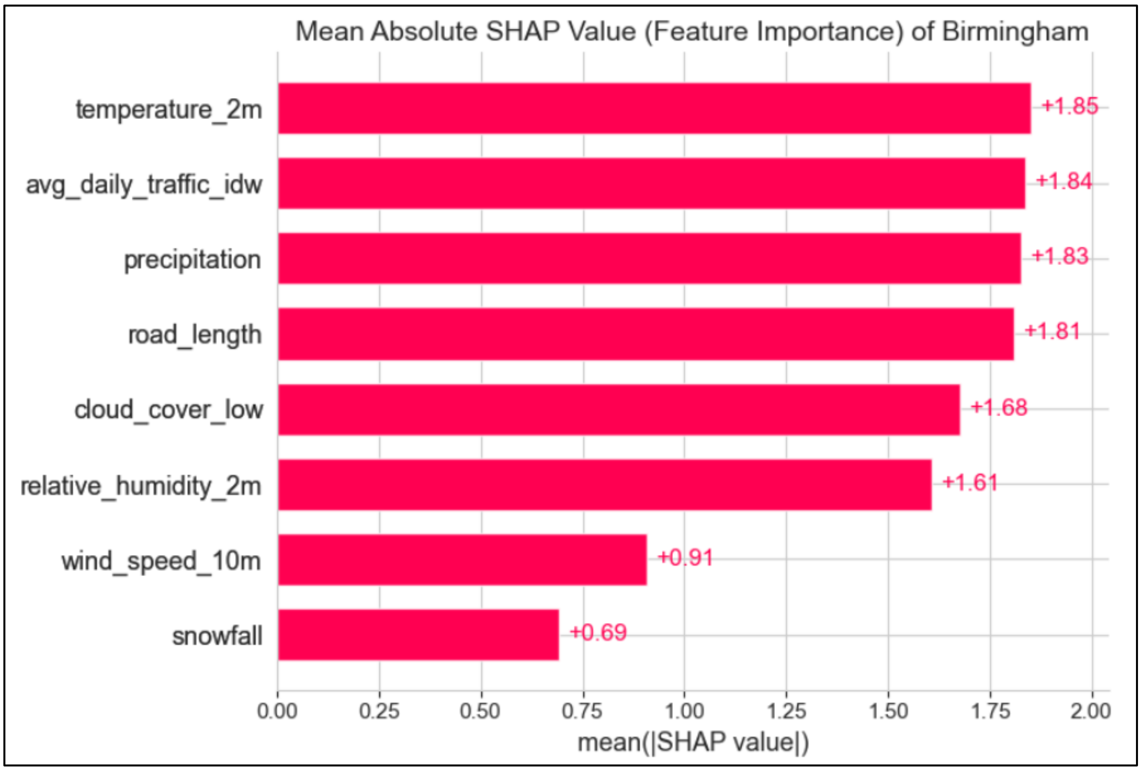

Figure 10: SHAP Value Feature Importance of Birmingham

The model is stronger than Manchester, possibly due to the data availability (R-squared = 0.550). However, it struggles with precise prediction, which aligns with the conclusion of spatial heterogeneity. Temperature and traffic volume have the strongest influence. Similar to Manchester, road length, wet conditions, and low visibility also influence the crash risk. However, unlike Manchester, weather conditions like humidity, wind speed, and snowfall have a much larger risk for traffic conditions.

The SHAP values strengthen the conclusion that the risk in Manchester is structurally driven, whereas in Birmingham it is behavioural and environmentally driven. Traffic accident data is noisy and influenced by several unobserved and complex factors. Also, the spatial and environmental data are merged, which introduces much variability. This explains the performance of the model. Since this is an urban crash risk and the goal is exploratory rather than predictive, the results achieved are acceptable. This complements the use of XGBoost for predictive modelling, since it outperforms the above models in structured and imbalanced data.

It must be noted that the GWR model for Manchester, though achieving low R-squared has better MAE and RMSE values. It is good for prediction whereas the Birmingham model is better for exploratory analysis. For this study, it can be concluded that the GWR model for Manchester performs better after analysis all the three metrics.

4.4 Objective 4 - Develop an XGBOOST model

An XGBoost classification model was developed to predict high-risk roads based on spatial, crash risk, and environmental variables. The target variable is ‘high_risk’, which is defined as 25% of the riskiest roads in terms of crashes per km. Binary variables are created: high_risk (1) where the road is among the 25% of the busiest roads and not (0).

The following features were selected for modelling:

Spatial features: *road_length, avg_daily_traffic_idw*

Weather features: *temperature_2m, precipitation, wind_speed_10m, relative_humidity_2m*

The data is split into training (75%) and test (25%) sets. This is the common practice in almost all the studies.

The initial baseline model missed several high-risk roads. Traffic crashes are rare occurrences that make class imbalance a significant limitation in crash modelling. This is handled through scale_pos_weight to give more weight to high-risk roads. Through grid search, the model is tuned with high recall and low precision.

Finally, extended features were added to the model, and with grid search, an optimal predictive model was achieved.

Spatial features*: road_length, avg_daily_traffic_idw, crashes_per_1000_vehicles_imputed*

Weather features: *temperature_2m, precipitation, wind_speed_10m, relative_humidity_2m, snowfall, cloud_cover_low*

Crash risk features: *avg_severity*

For cross-city analysis, two binary classifiers – one for Manchester and one for Birmingham – are trained and evaluated within their own city. It is cross-tested in the other city to analyse transferability.

Finally, the overall optimal predictive model with extended features is trained and tested on each of the cities separately.

4.4.1 Overall Predictive Model

Metric	Not high risk (0)	High risk (1)	Weighted average
Precision	0.93	0.82	0.90
Recall	0.94	0.78	0.90
F1-score	0.94	0.80	0.90
Support	179	60	239
Accuracy			0.90

Table 12: Performance of Overall XGBoost model

The model performs strongly with good performance across all the metrics. There is a strong balance between identifying high-risk roads and avoiding false positives.

4.4.2 Manchester Model

Metric	Not high risk (0)	High risk (1)	Weighted average
Precision	0.91	0.33	0.85
Recall	0.94	0.25	0.86
F1-Score	0.93	0.29	0.86
Support	66	8	74
Accuracy			0.86

Table 13: Performance of XGBoost model of Manchester

The model tested and trained in Manchester struggled to identify high-risk roads. This is due to a limited and imbalanced dataset that resulted in high accuracy but low precision.

4.4.3 Birmingham Model

Metric	Not high risk (0)	High risk (1)	Weighted Average
Precision	0.86	0.85	0.86
Recall	0.95	0.67	0.86
F1-Score	0.90	0.75	0.86
Support	113	52	165
Accuracy			0.86

Table 14: Performance of XGBoost model of Birmingham

Birmingham has a richer dataset compared to Manchester, with larger data and more high-risk roads. This has resulted in good precision and recall simultaneously.

4.4.4 Cross-Evaluation

Metric	Not high risk (0)	Weighted Average
Precision	0.74	0.73
Recall	0.95	0.74
F1-Score	0.83	0.70
Support	113	165
Accuracy		0.74

Table 15: Performance of XGBoost model of Manchester tested on Birmingham

Metric	Not high risk (0)	Weighted Average
Precision	0.97	0.92
Recall	0.91	0.89
F1-Score	0.94	0.90
Support	66	74
Accuracy		0.89

Table 16: Performance of XGBoost model of Birmingham tested on Manchester

The Manchester model struggles and fails to transfer to the Birmingham model. However, the Birmingham model generalises very well to the Manchester model, outperforming its own local model. This asserts the importance of a rich dataset. Birmingham's larger and balanced data perform strongly and even transfer well to models with sparse and imbalanced data.

4.4.5 City Evaluation

Metric	Value
Accuracy	98%
Precision (Class 1)	90%
Recall (Class 1)	93%
F1-Score (Class 1)	92%
ROC AUC	0.994
Confusion Matrix	[[262, 3], [2, 28]]

Table 17: Performance of Overall XGBoost model tested on Manchester

Metric	Value
Accuracy	97%
Precision (Class 1)	97%
Recall (Class 1)	95%
F1-Score (Class 1)	96%
ROC AUC	0.994
Confusion Matrix	[[443, 7], [11, 198]]

Table 18: Performance of Overall XGBoost model tested on Birmingham

The model demonstrates high generalisability across both cities with consistent metrics. Manchester has slightly better performance than Birmingham. These results validate the model's robustness and support its use for urban crash risk prediction in diverse environments.

On the whole, Birmingham performs better than the Manchester. This can be due nature of datasets. Birmingham has better dataset as it is larger with more high-risk roads than Manchester. Manchester has severely imbalanced dataset. There are only 8/74 high-risk roads in Manchester whereas there are 52/165 in Birmingham. This has led to better predictions in Birmingham than Manchester. This also leads to Birmingham model strong enough for transferability. Hence, with this framework, cities like Manchester that have sparse datasets can depend on cities like Birmingham that have much richer datasets for better predictions. The most common limitation in almost all the urban traffic studies is the presence of sparse data, which can be handled through this framework.

5. Policy Recommendations

Manchester:

Manchester requires targeted, structure-driven interventions and corridor-based strategies. Focus should be on major transport corridors - Oxford/Wilmslow Road, Stockport Corridor, Salford, and Ashton Old Road. These roads need strict traffic laws and regulation. This can be through reviewing speed limits, pedestrian facilities, posting warning signs and crash barriers since most of the accidents seems to be on highways and university roads. There is a need to design elements at intersections in peripheral areas that have high density to reduce accidents. This can be through proper signals or bikeway setbacks. Wilmslow Road needs significant measures and needs to be inspected as it has been a dangerous place for persistent accidents. There is a minimal seasonal but higher temporal pattern. There should be more traffic authorities present on Friday, especially at major public transport and arterial roads. It could be better if public is informed to be more careful, especially on Fridays for safer driving.

Manchester suffers due to structural factors like road length. Longer and complex roads and roads at intersections needs to be monitored. Crash data here could be furthered analysed with respect to road structure. This is especially important due to its dense and interconnected road network with findings that there is more traffic volume at key intersections. Wet conditions and low visibility seems to have an influence here. The use of headlights needs to be encouraged. There needs to be better lighting, advanced markings, beacons, and studs to improve visibility. High Friction Surfacing (HFS) is a targeted strategy for skid resistance, especially for road networks like that of Manchester. It has been proven by research to reduce wet weather accidents in the UK and thus could be implemented here.

Birmingham:

Birmingham requires systematic, city-wide strategies that focus on behavioural and environmental factors. It has much higher number of accidents than Manchester with much persistent hotspots that is spread across the city. Major arterial corridors - A34 (Stratford Road), A38 (Bristol/Kingsbury Road), A45 (Coventry Road), and A4540 (Ringoad) needs to be inspected through further audits. Seasonal pattern can be seen here and hence there is a need for more measures during November and December. There is a high weekend risk especially in Sundays that needs special attention through patrols. Bordsley Park Road and Haden Circus in the Birmingham Ring Road needs to be inspected as it is found to be the most dangerous areas. After this, attention should be given to Bristol Street, Stratford Road and Nechells. Commercial and industrial areas are shown to be risker and needs specific interventions.

Microclimate seems to have more effect here, which is consistent with the seasonal pattern. Weather conditions like humidity, wind speed, and snowfall have much more effect here. Microclimate-responsive plans need to be implemented with measures like snow clearance and traction control in areas with high snowfall sensitivity. There need to be wind barriers and humidity-risk materials in high-risk areas since these factors seems to differ by location in Manchester. Since behavioural factors play a much larger role here, public needs to be educated. Public awareness campaigns need to be conducted and special programmes in schools needs to be included. Safe driving habits like wearing seat belts, avoiding drunk driving, and paying attention to the speed needs to be encouraged. Also, there needs to be authorities to check for these at major road junctions.

6. Conclusion

In a systematic review by Noushin Behboudi et al. (2024), several directions for future research were given: spatiotemporal modelling, integration of diverse sources, interpretability, and transfer learning for cross-regions. All these have been addressed in this study that makes it a valuable contribution in the urban traffic safety field. Top performing models identified: random forest, XGBOOST, and SHAP were used for feature importance, prediction, and explainability. The four major objectives have been effectively addressed and the findings arrived from these have been used for providing policy recommendations.

6.1 Limitations

Though this dissertation addresses several critical gaps in the field and provides a foundation for future comparative traffic risk studies, there are limitations that needs to be considered. All the data sources used has its own challenges like underreporting, limited accuracy, and estimations that may affect the extent of analysis conducted. The study is focused on two urban cities in the same country, so it may not be generalisable to rural areas or cities in other countries. However, the framework developed in this study can be implemented after adjusting for the local variables.

Behavioural, demographic, and socioeconomic factors are excluded from this study, which limits the depth of the analysis. Though multiple data sources are integrated, there are several other potential influencing factors that are needed for complete crash risk modelling. Also, the machine learning models used may not establish causal relationships between risk factors and crashes. This has been addressed by combining the results with domain knowledge. This includes measures like road network metrics for traffic flow, feature enginering variables instead of using raw counts, using SHAP for interpretation and providing possible real-world explanations for the results.

Additionally, this study is conducted for the years 2019-2023. The data during 2020 and 2021 may be skewed due to the Covid pandemic. However, wherever possible this has been duly addressed with reasons in the analysis.

6.2 Future Research

Following a similar approach, socioeconomic, psychological, cultural, and behavioural factors need to be included for such comparative analysis between urban cities. These complex factors can be further analysed through cognitive studies. Similarly, multiple cities can be further studied. New models that are currently being developed, like neural networks or deep learning models, could be used in the future.

Rather than relying on open and government datasets, real time data of drivers, vehicles, and environment can be used for more accurate predictions. As mentioned, there are several factors that influence traffic accidents and only the key ones are used. Much other variables can be experimented with use the same framework.

There are more emerging technologies in development. These include artificial intelligence, autonomous vehicles, smart traffic signals, V2X technology, and much more than can be explore through similar comparative framework.

REFERENCES

Mesquitela, J., Elvas, L.B., Ferreira, J.C. and Nunes, L. 2022. Data Analytics Process over Road Accidents Data—A Case Study of Lisbon City. *ISPRS international journal of geo-information*. 11(2), p.143

World Health Organization. 2023. *Global Status Report on Road Safety 2023*. Geneva: WHO. Accessible from:

https://www.who.int/teams/social-determinants-of-health/safety-and-mobility/global-status-report-on-road-safety-2023

Srikanth, L. and Srikanth, I. 2020. A Case Study on Kernel Density Estimation and Hotspot Analysis Methods in Traffic Safety Management *In*: *International Conference on Communication Systems and Networks (Online)*. IEEE, pp.99–104

Abdulhafedh, A. 2017. Road Crash Prediction Models: Different Statistical Modeling Approaches. In: *Journal of Transportation Technologies*, 7(2). Scientific Research Publishing, pp.190–205

Li, H. and Chen, L. 2025. Traffic accident risk prediction based on deep learning and spatiotemporal features of vehicle trajectories. *PloS one*. 20(5), p.e0320656

Alsahfi, T. 2024. Spatial and Temporal Analysis of Road Traffic Accidents in Major Californian Cities Using a Geographic Information System. *ISPRS international journal of geo-information*. 13(5), p.157

Xia, H., Liu, R., Zhou, W. and Luo, W. 2024. Modeling the Causes of Urban Traffic Crashes: Accounting for Spatiotemporal Instability in Cities. *Sustainability*. 16(20), p.9102

Vivas Pacheco, H., Rodríguez-Mariaca, D., Jaramillo, C., Fandiño-Losada, A. and Gutiérrez-Martínez, M.I. 2023. Traffic Fatalities and Urban Infrastructure: A Spatial Variability Study Using

Geographically Weighted Poisson Regression Applied in Cali (Colombia). *Safety (Basel)*. 9(2), p.34

Musingura, C., Lee, G., Ahn, Y. and Kim, K., 2023. Mitigating Road Traffic Crashes in Urban Environments: A Case Study and Literature Review-based Approach. *Authorea Preprints*

Abdalazeem, M. and Oke, J. 2025. Roadway Crash Typology of Census Tracts Enables Targeted Interventions via Interpretable Machine Learning. *Data science for Transportation*. 7(2)

Alpalhão, N., Sarmento, P., Jardim, B. and de Castro Neto, M., 2025. Assessing the risk of traffic accidents in lisbon using a gradient boosting algorithm with a hybrid classification/regression approach. *Transportation Research Interdisciplinary Perspectives*, *32*, p.101495

Jin, J., Liu, P., Huang, H. and Dong, Y. 2024. Analyzing urban traffic crash patterns through spatio-temporal data: A city-level study using a sparse non-negative matrix factorization model with spatial constraints approach. *Applied geography (Sevenoaks)*. 172.

He, M., Meng, G., Wu, X., Han, X. and Fan, J., 2025. Road Traffic Accident Prediction Based on Multi-Source Data–A Systematic Review. *Promet-Traffic&Transportation*, *37*(2), pp.499-522.

Skaug, L., Nojoumian, M., Dang, N. and Yap, A. 2025. Road Crash Analysis and Modeling: A Systematic Review of Methods, Data, and Emerging Technologies. *Applied sciences*. 15(13), p.7115.

International Transport Forum, 2018. *ITF Research Reports Safer City Streets Global Benchmarking for Urban Road Safety*. OECD Publishing.

Sun, D., Ai, Y., Sun, Y. and Zhao, L. 2020. A highway crash risk assessment method based on traffic safety state division. *PloS one*. 15(1), p.e0227609.

Butt, M.S. and Shafique, M.A. 2025. A literature review: AI models for road safety for prediction of crash frequency and severity. *Discover Civil Engineering*. 2(1).

Merlin, L.A., Guerra, E. and Dumbaugh, E. 2020. Crash risk, crash exposure, and the built environment: A conceptual review. *Accident analysis and prevention*. 134.

Pei, Y., Babcock, B., Jin, L. and Luk, W.S., 2012. *Towards a taxonomy of web search result diversification*. Proceedings of the 20th International Conference on World Wide Web, pp. 631–640.

Greenshields, B.D., 1935. A study in highway capacity. *Highway Research Board Proc., 1935*, pp.448-477.

Haghani, M., Coughlan, M., Crabb, B., Dierickx, A., Feliciani, C., van Gelder, R., Geoerg, P., Hocaoglu, N., Laws, S., Lovreglio, R. and Miles, Z., 2023. Contemporary challenges in crowd safety research and practice, and a roadmap for the future: The Swiss Cheese Model of Crowd Safety and the need for a Vision Zero target.

Giménez-Santana, A., Medina-Sarmiento, J.E. and Miró-Llinares, F. 2018. Risk terrain modeling for road safety: identifying crash-related environmental factors in the province of Cádiz, Spain. *European journal on criminal policy and research*. 24(4), pp.451–467.

Dumitrascu, D.I., 2024. Influence of Road Infrastructure Design over the Traffic Accidents: A Simulated Case Study. *Infrastructures*, *9*(9), p.154.

Dong, Z. and Guo, C. 2021. A Literature Review of Spatio-temporal Data Analysis. *Journal of physics. Conference series*. 1792(1).

Koutsaki, E., Vardakis, G. and Papadakis, N., 2025. Event Prediction Using Spatial–Temporal Data for a Predictive Traffic Accident Approach Through Categorical Logic. *Data*, *10*(6), p.85.

Choudhary, A., Mishra, V., Garg, R.D. and Jain, S.S., 2025. Spatio–temporal analysis of traffic crash hotspots-an application of GIS-based technique in road safety. *Applied Geomatics*, pp.1-18.

Yuan, J., Zheng, Y. and Xie, X. 2012. Discovering regions of different functions in a city using human mobility and POIs *In*: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, pp.186–194.

Li, H. and Chen, L. 2025. Traffic accident risk prediction based on deep learning and spatiotemporal features of vehicle trajectories. *PloS one*. 20(5), p.e0320656.

Shi, Z. and Pun-Cheng, L.S.C. 2019. Spatiotemporal Data Clustering: A Survey of Methods. *ISPRS international journal of geo-information*. 8(3), p.112.

Chaudhuri, S., Juan, P. and Mateu, J. 2023. Spatio-temporal modeling of traffic accidents incidence on urban road networks based on an explicit network triangulation. *Journal of applied statistics*. 50(16), pp.3229–3250.

Saini, A., Gauba, N., Chawla, H. and Ali, J.A.B.I.R., 2021. Road Accidents Analysis Using Comparative Study & Application of Machine Learning Algorithms. *WSEAS Trans Comput Res*, *9*, pp.78-86.

Yalamanchili, S., 2024. Data linkage in road safety: bridging the divide to support better health outcomes.

Çepni, M.S. and Alp, E., 2017. Erken Tarihte Yapılan İmar Uygulamalarının Kentleşmeye Etkisi: Körfez-Kocaeli Örneği. 16. *Türkiye Harita Bilimsel Ve Teknik Kurultayı, Ankara*.

Furlong, J., Fevyer, D., Armstrong, B., Edwards, P., Aldred, R. and Goodman, A. 2025. Low Traffic Neighbourhoods in London reduce road traffic injuries: a controlled before-and-after analysis (2012–2024). *Injury prevention*., p.ip–2024–045571.

Wahab, L. and Jiang, H. 2019. A comparative study on machine learning based algorithms for prediction of motorcycle crash severity. *PloS one*. 14(4), p.e0214966.

Li, F., Li, Y. and Rogerson, L.E. 2025. London Blue Light Collaboration Evaluation: A Comparative Analysis of Spatio temporal Patterns on Emergency Services by London Ambulance Service and London Fire Brigade.

Kumar, K. S., Sai, P. V., Kiran, P., & Sreekanth, S. (2025). Implementing machine learning algorithms for classifying the data – Random Forest, XGBoost, LSTM, Hybrid Algorithm. *Journal of Emerging Technologies and Innovative Research (JETIR)*, 12(4).

Jamal, A., Zahid, M., Tauhidur Rahman, M., Al-Ahmadi, H.M., Almoshaogeh, M., Farooq, D. and Ahmad, M. 2021. Injury severity prediction of traffic crashes with ensemble machine learning techniques: a comparative study. *International journal of injury control and safety promotion*. 28(4), pp.408–427.

Zhang, J., Li, Z., Pu, Z. and Xu, C. 2018. Comparing Prediction Performance for Crash Injury Severity Among Various Machine Learning and Statistical Methods. *IEEE access*. 6, pp.60079–60087.

Chen, F., Liu, X.Q., Yang, J.J., Liu, X.K., Ma, J.H., Chen, J. and Xiao, H.Y. 2025. Traffic accident severity prediction based on an enhanced MSCPO-XGBoost hybrid model. *Scientific reports*. 15(1).

Panda, C., Mishra, A.K., Dash, A.K. and Nawab, H. 2023. Predicting and explaining severity of road accident using artificial intelligence techniques, SHAP and feature analysis. *International journal of crashworthiness*. 28(2), pp.186–201.

Tang, J., Huang, Y., Liu, D., Xiong, L. and Bu, R. 2025. Research on Traffic Accident Severity Level Prediction Model Based on Improved Machine Learning. *Systems (Basel)*. 13(1), p.31.

Chen, Y., Tian, Y., Ouyang, Z. and Zhu, J. 2025. Influence of road environmental factors on traffic accidents involving vulnerable road users through negative binomial models. *PloS one*. 20(2), p.e0317601.

Ren, H., Song, Y., Wang, J., Hu, Y. and Lei, J., 2018, November. A deep learning approach to the citywide traffic accident risk prediction. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)* (pp. 3346-3351). IEEE.

Liang, Y., Yuan, H., Wang, Z., Wan, Z., Liu, T., Wu, B., Chen, S. and Tang, X. 2024. Nonlinear effects of traffic statuses and road geometries on highway traffic accident severity: A machine learning approach. *PloS one*. 19(11), p.e0314133.

Carrodano, C. 2024. Data-driven risk analysis of nonlinear factor interactions in road safety using Bayesian networks. *Scientific reports*. 14(1).

Stiles, J. and Miller, H.J. 2024. The built environment and the determination of fault in urban pedestrian crashes: Toward a systems-oriented crash investigation. *Journal of transport and land use*. 17(1), pp.97–113.

Gedamu, W.T., Plank-Wiedenbeck, U. and Wodajo, B.T., 2025. Spatio-temporal analysis of road traffic crashes by severity. *Transportation Engineering*, *20*, p.100327.

Shikder, Md.F.H., Tang, Y. and Emami Javanmard, M. 2025. Time-Series Forecasting for Peak Hour Traffic Accidents. *IEEE open journal of intelligent transportation systems*. 6, pp.883–897.

Harith, S.H., Mahmud, N. and Doulatabadi, M., 2019, March. Environmental factor and road accident: a review paper. In *Proceedings of the international conference on industrial engineering and operations management* (pp. 3409-3418).

Mohammed, O., 2025. Understanding the Impact of Driver Behavior on Traffic Safety: A Comprehensive Review of Behavioral, Technological, and Environmental Factors. *Al-Rafidain Journal of Engineering Sciences*, pp.626-642.

Durap, A., 2025. Interpretable machine learning for coastal wind prediction: Integrating SHAP analysis and seasonal trends. *Journal of Coastal Conservation*, *29*(3), p.24.

City Population (n.d.) *United Kingdom: Urban Areas*. Accessible from:

https://www.citypopulation.de/en/uk/cities/ua/

Greater Manchester Combined Authority (n.d.) *Vision Zero*. Accessible from:

https://www.greatermanchester-ca.gov.uk/what-we-do/transport/vision-zero/ (Accessed: 2 September 2025).

Birmingham City Council (2025) *Road Harm Reduction Strategy and Action Plan*. Accessible from:

https://www.birmingham.gov.uk/info/50348/transport_plan_and_policies/3048/road_harm_reduction_strategy_and_action_plan

Huang, H., Zhou, H., Wang, J., Chang, F. and Ma, M., 2017. A multivariate spatial model of crash frequency by transportation modes for urban intersections. *Analytic methods in accident research*, *14*, pp.10-21.

Doudaran, M.S., 2020. Traffic Injury Prevention Techniques Using STATS19 Road Safety Data: A Model-comparison Approach.

Lovelace, R., 2020. Reproducible road safety research with R: A practical introduction.

Department for Transport (2024) *Reported road casualties Great Britain: annual report 2023*. London: GOV.UK. Accessible from:

https://www.gov.uk/government/statistics/reported-road-casualties-great-britain-annual-report-2023/reported-road-casualties-great-britain-annual-report-2023

Chang, V., Xu, Q.A., Hall, K., Oluwaseyi, O.T. and Luo, J. 2023. Comprehensive analysis of UK AADF traffic dataset set within four geographical regions of England. *Expert systems*. 40(10).

Boeing, G. 2017. OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*. **65**, pp.126–139.

Neelima, N.V., Wu, S., Shriyam, S. and Wei, Y., 2024, December. Road Network and its Impact on Urban Socio-ecological Systems: A Network Perspective. In *International Conference on Complex Networks and Their Applications* (pp. 425-434). Cham: Springer Nature Switzerland.

Behboudi, N., Moosavi, S. and Ramnath, R. 2024. Recent Advances in Traffic Accident Analysis and Prediction: A Comprehensive Review of Machine Learning Techniques.