Deciphering Car Crash Dynamics in Greater Melbourne: A Multi-Model Machine Learning and Geospatial Analysis

by
Christopher Johnson

Supervisor: Chayn Sun

Capstone Project Report

Abstract

In the ever-evolving landscape of data-driven methodologies addressing car crash patterns, a holistic analysis remains critical to decode the complex nuances of this phenomenon. This study bridges this knowledge gap with a robust examination of car crash occurrence dynamics and the influencing variables in the Greater Melbourne area, Australia. We employed a comprehensive multi-model machine learning and geospatial analytics approach, unveiling the complicated interactions intrinsic to vehicular incidents. By harnessing RF with SHAP, GLR, and GWR, our research not only highlighted pivotal contributing elements but also enriched our findings by capturing often overlooked complexities. Using the Random Forest model, essential factors were emphasised, and with the aid of SHAP (Shapley Additive Explanations), we accessed the interaction of these factors. To complement our methodology, we incorporate hexagonalised geographic units, refining the granularity of crash density evaluations. In our multi-model study of car crash dynamics in Greater Melbourne, road geometry emerged as a key factor, with intersections showing a significant positive correlation with crashes. The average land surface temperature had variable significance across scales. Socio-economically, regions with a higher proportion of childless populations were identified as more prone to accidents. Public transit usage displayed a strong positive association with crashes, especially in densely populated areas. The convergence of insights from both Generalized Linear Regression and Random Forest's SHAP values offered a comprehensive understanding of underlying patterns, pinpointing high-risk zones and influential determinants. These findings offer pivotal insights for targeted safety interventions in Greater Melbourne

1. Introduction

Vehicle crashes create a significant global concern with impacts on public health, economic stability, and societal well-being. Millions of people are affected by vehicular accidents, with approximately 1.3 million deaths and 50 million injuries each year globally (World Health Organization 2018). The human costs of these accidents, coupled with the enormous economic burden of medical care, property damage, and productivity loss, highlight the need for effective preventive strategies.

To devise these strategies, it is essential to understand the complex causes and influencing factors behind vehicular accidents. Urban areas, in particular, offer a complex set of interplaying variables such as road infrastructure, environmental conditions, socio-economic attributes, and demographic characteristics, all of which can influence accident rates. Among urban regions, Greater Melbourne stands out due to its intricate urban environment and diversity in socio-economic and environmental aspects. This region, therefore, provides a unique opportunity to delve deep into the multitude of factors that contribute to road safety issues.

Machine learning and geospatial analytics have emerged as potent tools for unravelling such complexities, allowing for the identification of high-risk areas and pinpointing the significant determinants of road accidents. By employing these advanced analytical methods, it is possible to generate predictive models and understand patterns that might otherwise be obscured in traditional analytical frameworks.

This study aims to utilise a multi-model approach to improve robustness of findings and facilitate a more in-depth analysis of contributing factors. Further strengthening our methodology is our choice of a hexagonalised dataset over traditional statistical area boundaries. This decision emerges from the understanding that hexagonal datasets mitigate challenges presented by the Modifiable Areal Unit Problem (MAUP). Furthermore, by combining various datasets, from road attributes to socio-economic indicators, we seek to present a comprehensive view of the accident landscape in this urban area. The ultimate goal is to offer actionable insights and recommendations that can guide urban planners, policymakers, and stakeholders in creating safer road environments and implementing preventive measures more effectively.

2. Literature Review

2.1Factors Affecting Road Crashes

Environmental factors play a pivotal role in vehicle road crashes due to the direct and indirect impacts they have on drivers' behaviour and perception. For an example a significant factor that has been identified is the higher land-use density, represented by the urbanity level, which has been found to negatively impact traffic safety. In a comprehensive study conducted by Asadi et al. (2022), it was revealed that areas characterized by higher land-use density tend to experience a decrease in traffic safety levels. Dumbaugh and Zhang (Citation2013) find that commercial strips and big box stores are related to crashes by older adults.

Additional important factors to consider are socioeconomics and demographics. Such factors may directly affect behavioural patterns, lifestyle patterns, etc. The study by Al-Mistarehi et al. (2022) highlighted that younger drivers within the age range of 18-36 years were more likely to be involved in traffic crashes, indicating a correlation between driver age and the likelihood of crashes. Furthermore, a paper by Tavris, Kuhn & Layde (2001) studying the patterns in gender and motor vehicle crashes found a relation between males and vehicle crashes. It was suggested that males were more frequently involved in loss of control type crashes. Therefor gender may play a role in contributing towards total vehicle crashes.

Road geometry plays a critical role in influencing the occurrence of crashes because it directly affects the movement, behaviour, and interaction of road users. Wang and Zhang's (2017) findings indicated a positive relationship between speed limits and crash severity. The study concluded that an increase in speed limits led to a subsequent rise in the likelihood of severe crashes. Similar findings by Liu X and Xia J (2015) identified speed limits over 100 km/h to be significant. Studies such as Al-Mistarehi et al. (2022) have identified similar relationships between speed limits and crash severity. Wang and Zhang (2017) further emphasized the role of road alignment in influencing crash outcomes. They found that crashes on curved roads were more likely to be severe than those on straight roads, underscoring the risks of curved alignments. Moreover, their research pinpointed location as a significant determinant of crash severity. Specifically, crashes at intersections were generally less severe than those at non-intersection locations. Wang and Zhang (2017) also highlighted the varying effects of different road function classes on crash severity. Their findings indicated that rural roadways, particularly rural major arterials, had a higher likelihood of severe crashes compared to urban roads.

2.2Related Work and Research Gaps

Numerous studies align with our research, leveraging machine learning to investigate spatial patterns and relationships in road crashes. Liu X and Xia J (2015) primary objective of their study was to discern the spatial patterns of fatal single vehicle crashes (FSVC), particularly in Western Australia, and reveal the interaction between FSVC and possible determinants. The analysis incorporated two broad categories of factors: Road conditions and environmental factors. A Geographic Weighted Regression model was employed to assist in pinpointing the correlations between the discussed factors and FSVC hotspots. A pivotal finding of the study was the variance in the impact of these significant risk factors across rural and metropolitan areas.

An additional related study by Lee D et al. (2018) examined car crashes in the Central Ohio Region between 2006 and 2011, emphasizing the correlation between the surrounding socio-economic and built environments. Recognizing potential spatial autocorrelation influences, the study also integrates spatial econometric models for accuracy. Conclusively, the research offers insights on transportation safety policy, based on its findings.

A study by Rhee et al. (2016) conducted a comprehensive spatial analysis of traffic crashes in Seoul using a multi-model approach. Their study aggregated data at the traffic analysis zone (TAZ) level and employed a combination of spatial error and lag models, Ordinary Least Squares (OLS) Regression, and Geographically Weighted Regression (GWR) to examine the factors influencing vehicle crashes.

Recent research, including a study by Elyassami, Hamid, & Habuza (2020), utilized random forest models to explore critical risk factors associated with road accidents. Their objective was to develop a model that precisely predicts the severity of driver injuries. The study yielded promising results, effectively highlighting that road design flaws were among the most significant contributing factors.

To the best of our knowledge, several research gaps have emerged. A prevalent trend among many studies is the use of administrative boundaries for data aggregation, leading to issues related to the Modifiable Areal Unit Problem (MAUP), which we delve into later in this paper. In our approach, we adopt a hexagonal dataset for aggregation, aiming to mitigate these challenges and evaluate its feasibility. Moreover, we intend to broaden the range of features employed in past studies. For instance, while some previous research has primarily centred on road and environmental factors, our study seeks to integrate a more varied array of attributes, including additional socio-economic and demographic data. Furthermore, it's noteworthy that factors influencing crashes can differ considerably based on geographical location. Hence, findings from one country or state might not align with those from another. In our exploration, we've noticed a scarcity of in-depth spatial analyses of road crashes specifically within the Greater Melbourne region.

3. Methodology

3.1Study Area and Data Sources

The selected research area was the Greater Melbourne region, an area characterized by its urban complexity and diverse range of environmental, road, socio-economic, and demographic attributes. Such complexity underscores the need for sophisticated analytical methods, prompting the use of machine learning models like Random Forest and Geographically Weighted Regression. Moreover, the robust data landscape of Greater Melbourne makes it ideal for building well-rounded datasets.

Our primary dataset comprised crash records obtained from VicRoads, detailing accident locations spanning five years, from 2015 to 2020. Data was sourced from OpenStreetMap, encompassing the Melbourne road network and Points of Interest (POIs). In addition, satellite-derived metrics were incorporated, including the LST (Land Surface Temperature), NDBI (Normalized Difference Built-Up Index), and NDVI (Normalized Difference Vegetation Index). To supplement our analysis, we also incorporated 2020 census data, offering a comprehensive view of the socio-economic and demographic landscape of the region.

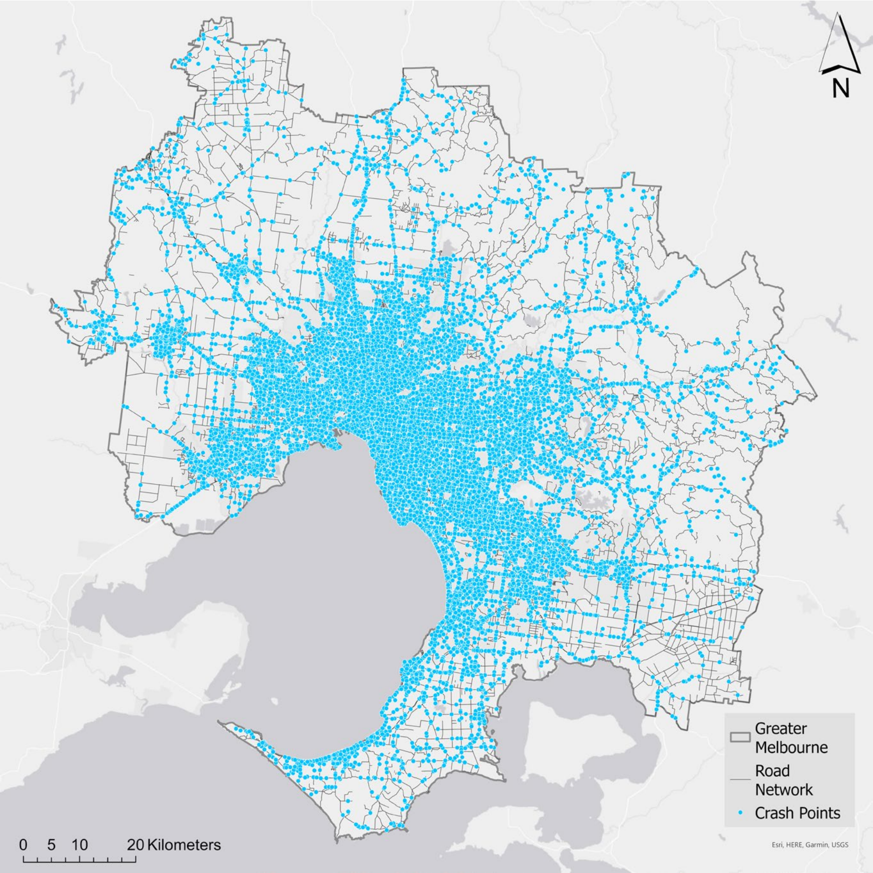


Figure 1: Study Area and Distribution of Crash Points

Table 1: Overview of Features used in the Study

Feature	**Description**	**Category**
Crash Count	Total crashes occurring in a hexagon	Independent Variable
Residential Road Length	Total length (m) of residential roads	Road Geometry
Other Road Length	Total length (m) of roads that are not residential or road links	
Link Road Length	Total length (m) of link roads such as ramps	
Total Nodes	Count of intersections between two or more streets	
Total Roundabouts	Count of intersections classified as roundabouts	
Avg. Roads at Intersections	Average number of roads that connect to an intersection	
Avg. Speed Limit	The average speed limit of a road network	
Avg. Road Curvature	Measured of average curvature	
Total Bridges	Total number of road segments classified as bridges	
Total Tunnels	Total number of road segments classified as tunnels	
Population Aged 65+ (%)	Ratio of population over the age of 65	Socioeconomics and Demographics
Male Population (%)	Ratio of male population	
Private Transit Usage (%)	Ratio of private transit usage	
Public Transit Usage (%)	Ratio of public transit usage	
Manager, admin, and prof workers (%)	Ratio of managers, administrative and professional workers	
Low Income Population (%)	Ratio of low-income population	
English Speaking Population (%)	Ratio of English-speaking population	
Indigenous Population (%)	Ratio of indigenous population	
Childless Population (%)	Ratio of population that do not have children	
Avg. Vegetation Index	Average vegetation coverage	Environment
Avg. Land Surface Temperature	Average land surface temperature	
Avg. Built-Up Index	Average measure of urban density	

Avg. Slope of Roads	Average slope of a road network	
Sum Amenity POIs	Count of amenity POIs such as cafes and hospitals	
Sum Tourism POIs	Count of tourism POIs such as hotels and attractions	
Sum Leisure POIs	Count of leisure POIs such as parks and stadiums	
Sum Shop POIs	Count of shop POIs such as supermarkets and clothing	

3.2Multi-Technique Machine Learning Approach

Our study utilised three machine learning models: Random Forest, Generalised Linear Regression (GLR), and Geographically Weighted Regression (GWR). These models were selected as they complement each other’s strengths and weaknesses. The Random Forest model was used to capture the complex relationships between road crashes. Random Forest excels at capturing non-linear relationships, which the other models, GLR and GWR, might not effectively handle. Furthermore, Random Forest's capability to manage a large number of features and its inherent robustness to collinearity makes it an ideal candidate to provide initial general and global insights into the relationships of the dataset. Our Random Forest model serves as an introduction to machine learning results and analysis.

While Random Forest is excellent at capturing complex relationships, it may lack interpretability. SHAP values are used to explain the model's predictions by quantifying the contribution of each feature to the prediction. This aids in understanding how different contributing factors influence the increase of car accidents. SHAP values can indicate whether a specific feature increased or decreased the occurrences of car crashes. This information is crucial for understanding the direction of influence of each contributing factor.

The Generalised Linear Regression (GLR) model was used to further refine selected features for the GWR model. This refinement involved selecting only statistically significant values with a p-value of less than 0.01. Employing the GLR as a preliminary feature selection step serves multiple purposes. Firstly, it aids in dimensionality reduction, ensuring the Geographically Weighted Regression (GWR) model isn't burdened with variables that might not contribute meaningful information. By focusing on variables that are statistically significant in the GLR model, we are targeting features that have a clear linear relationship with the outcome variable. This methodical approach has the advantage of enhancing the stability and interpretability of the subsequent GWR model.

The GWR model was employed to account for the spatial autocorrelation of road crashes. It recognizes that incidents in one location may be influenced by nearby areas and allows for the relationship between variables to vary across space. Furthermore, GWR provides localised parameter estimates, acknowledging that contributing factors may have different effects in different areas. This is essential for accurately modelling road safety, as the impact of factors can vary significantly from one region to another.

3.3Hexagonalisation

H3 is a geospatial indexing system developed by Uber Technologies. It subdivides the world into hexagonal cells, providing a unique identifier for each cell. This hierarchical grid system offers multiple resolutions, allowing users to select the desired granularity for their specific use case. Each higher resolution results in smaller hexagons, thereby providing more detailed spatial divisions.

This study uses multi-model approach to take advantage of the systems benefits. Firstly H3's hierarchical nature allows for multi-resolution analysis. Different resolutions can capture patterns at various spatial scales, from broad to granular. This flexibility is essential when examining relationships that may form differently at various scales. Therefor we can easily build multiple datasets at designated scales. The multi-model H3 approach also enhances actionability for stakeholders, tailoring features to the most approprirate resolution based on feature scale and actionability.

Additionally, one of the longstanding challenges in spatial analysis is the Modifiable Areal Unit Problem (MAUP), where the choice of spatial aggregation can lead to different results. A study by Álvaro B-R et al. (2019) underscores this challenge, finding that changing the type of BSU, such as transitioning from census tracts to hexagonal units, had a more pronounced impact on the estimated values of model parameters than merely adjusting the size or scale of the BSUs. Notably, their research indicated that hexagonal units, when combined with a conditional autoregressive model, performed optimally. While no system can fully eliminate the MAUP, the spatial consistency of H3 provides a fair ground for comparisons across regions, ensuring that anomalies or patterns are genuine and not artifacts of the mapping system. This offers a significant advantage over traditional administrative boundaries, which might introduce biases or obscure certain spatial patterns. Such consistency is especially vital when employing models like GWR, where discerning local spatial variations is paramount.

3.4Data Pre-Processing

Each dataset initially underwent a rigorous data cleaning process to ensure compatibility with machine learning algorithms. This process involved addressing null values and discarding irrelevant data. Additional features were derived from existing data such as calculating the road curvature from the OSM road network and calculating the slope of each road segment using the Victorian DEM.

We utilised the H3 Python library to aggregate our datasets into hexagonal grids. We created two distinct H3 datasets at resolutions 7 and 8. The dataset at resolution 8 features a cell size of roughly 0.7 km², representing the neighbourhood scale. Conversely, the resolution 7 dataset, with its more expansive cell size of about 5 km², is designed to encapsulate the local government scale.

With the VicRoads crash dataset, we computed the aggregate count of crashes within each H3 cell. This metric acts as our target variable for analysing correlations and relationships. Meanwhile, values from the census and satellite datasets were averaged within respective cells. As for the data sourced from OpenStreetMap (OSM), we quantified the total number of Points of Interest (POIs), computed the cumulative length of roads, and determined the average speed limit within each hexagon.

We trained a Random Forest model on both datasets, using 'crash count' as the target variable. Feature importance scores were visualized using a bar chart to highlight the most influential features. Additionally, we utilized the SHAP Python library to provide a comprehensive summary of the SHAP values for each feature. For clarity, we grouped and presented plots by feature categories, such as 'environmental' and 'socio-economic', to differentiate between related factors more clearly.

Next features were assigned to either the local government or neighbourhood scale based on the features scale (how much a feature varies across a given space) and the actionability. These decisions were purely based on domain knowledge and human intelligence. For an example the percentage of male population is a relatively stable feature over a small area and therefor a larger hexagonal resolution would be more appropriate to capture such variation. In the context of actionability, average speed limit was deemed suitable at the neighbourhood scale given both the frequent variations in speed limits across roads and to pin point specific problematic areas based on local coefficients. Relevant features consistent across both scales, like the Avg. Built-Up Index, were incorporated into both datasets. We believe that these features are integral to the overall relationships and that their exclusion would render the model incomplete.

Before applying linear regression models, it is essential to check our dataset for any potential multi-collinearity issues. Serious multi-collinearity between features can cause unstable estimates and consequently less reliable results. We set a commonly used threshold of 10. Features with a VIF > 10 were iteratively eliminated before re-training the model. We stopped this process until all features were below a VIF value of 10, indicating no serious multi-collinearity issues. The final step involved training the GLR to perform a check for statistically significant features. A commonly accepted threshold p value of 0.05 was used. Features with a value greater than this were deemed statistically insignificant and subsequently removed. Below are the final two datasets after the feature splitting and both VIF checks and statistical significance checks. Table 2 shows the final features used in each dataset for the GWR model.

Table 2: Features used for the GWR models

Resolution 7 Features	**Resolution 8 Features**
Other Road Length	Avg. Vegetation Index
Total Nodes	Avg. Land Surface Temperature
Male Population (%)	Avg. Built-Up Index
Childless Population (%)	Total Bridges
Avg. Vegetation Index	Total Tunnels
Avg. Land Surface Temperature	Residential Road Length
Avg. Built-Up Index	Link Road Length
Total Bridges	Total Roundabouts
Total Tunnels	Avg. Roads at Intersections
English Speaking Population (%)	Avg. Speed Limit
Sum Tourism POIs	Avg. Road Curvature
Sum Leisure POIs	

Utilising the refined datasets we trained a GWR model for the neighbourhood and local government datasets in R. We extracted the local feature coefficients and the local R2 results. To plot the maps a combination of ArcGIS Pro and Python were used.

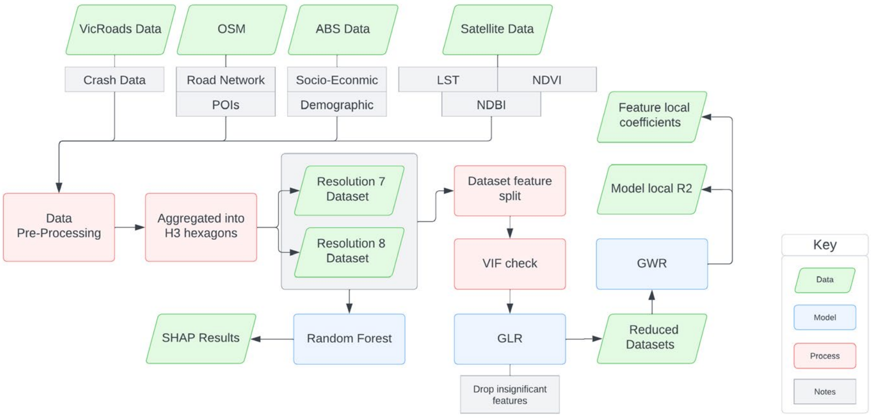
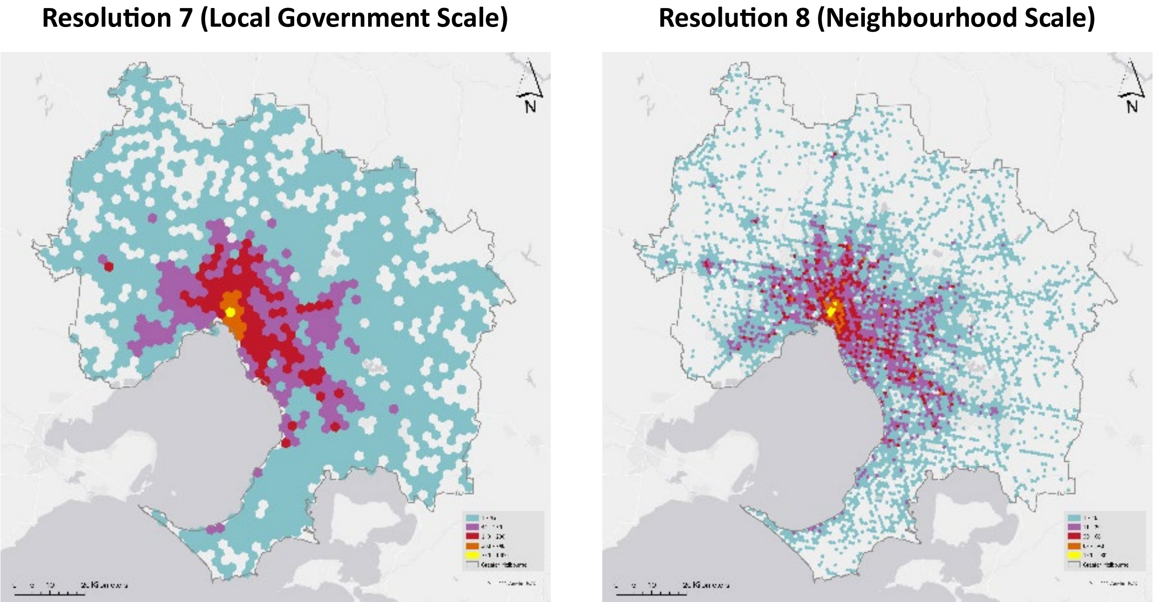


Figure 2: Methodological Framework of Study

4. Results

4.1 Crash Density


Resolution 7 (Local Government Scale)

Resolution 8 (Neighbourhood Scale)

Figure 3: Crash Distribution after Processing

Figure 3 highlights the distribution of total crashes. Notably, sub-urban areas hold higher counts of road crashes, which gradually decrease as we transition to the more remote outskirts of the Greater Melbourne region.

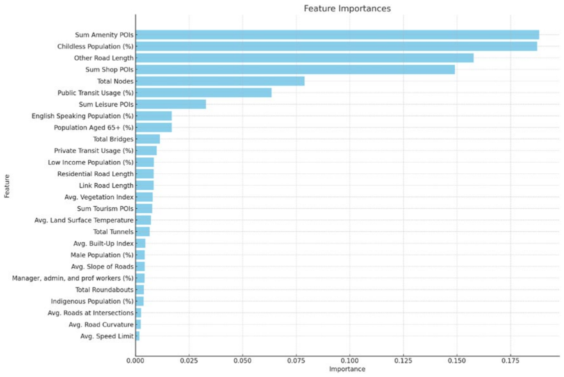
4.2 Random Forest Performance

Table 3: Random Forest Performance Metrics

Metric	Resolution 7	Resolution 8
R^2	0.88	0.75
Mean Absolute Error (MAE)	9.95	3.75

4.3 SHAP Summary: Global Impact Range and Distribution of Features

Resolution 7 (Local Government Scale)



Neighbourhood Scale Model

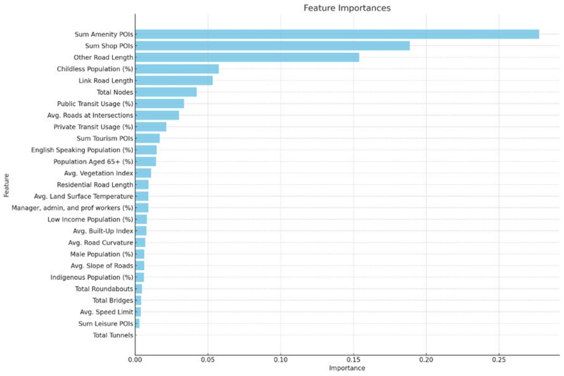
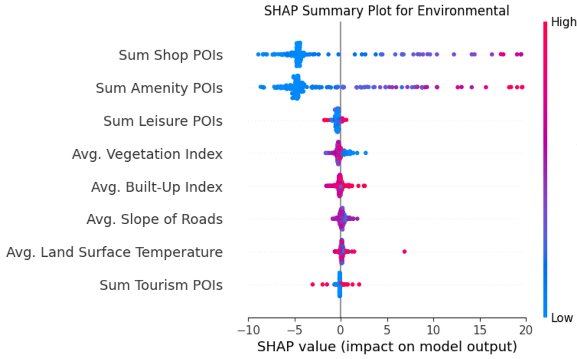
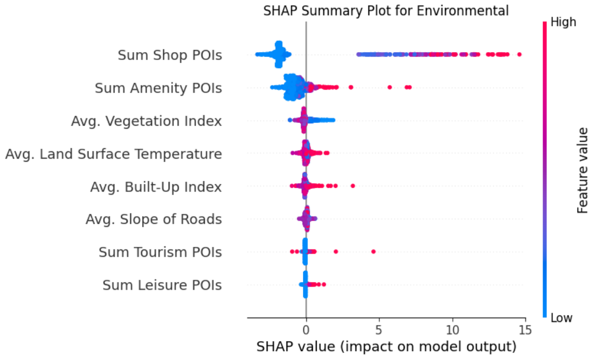


Figure 4: Random Forest Feature Importance

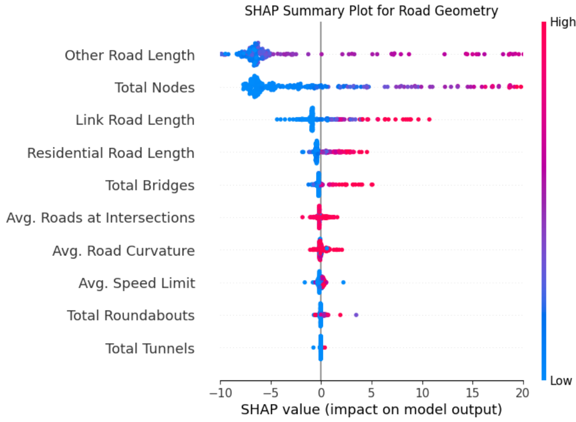
Resolution 7 (Local Government Scale) | **Resolution 8 (Neighbourhood Scale)**

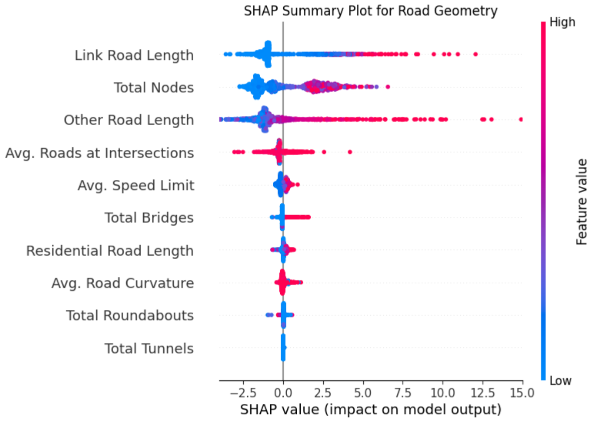
Environmental Factors



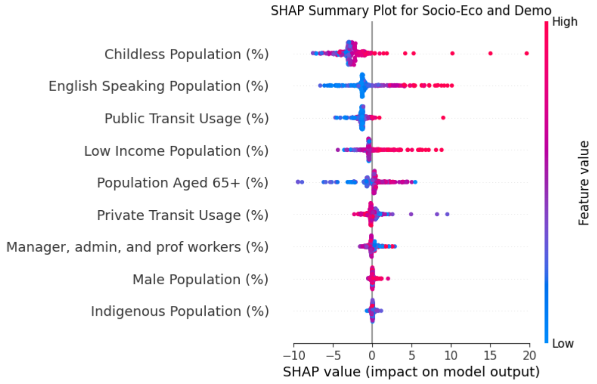


Road Geometry Factors





Socio-Economic and Demographic Factors



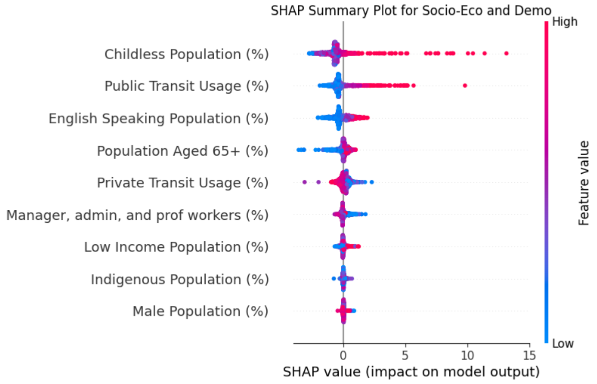


Figure 5: SHAP Summary Plots Categorised by Feature Type

The plot sorts feature by the sum of SHAP value magnitudes over all samples, which can show the distribution of the impacts each feature has on the model output. The colour represents the feature value (red high, blue low).

4.4 Generalised Linear Regression Results

Table 4: Resolution 8 (Neighbourhood Scale) GLR Results

R2 = 0.70

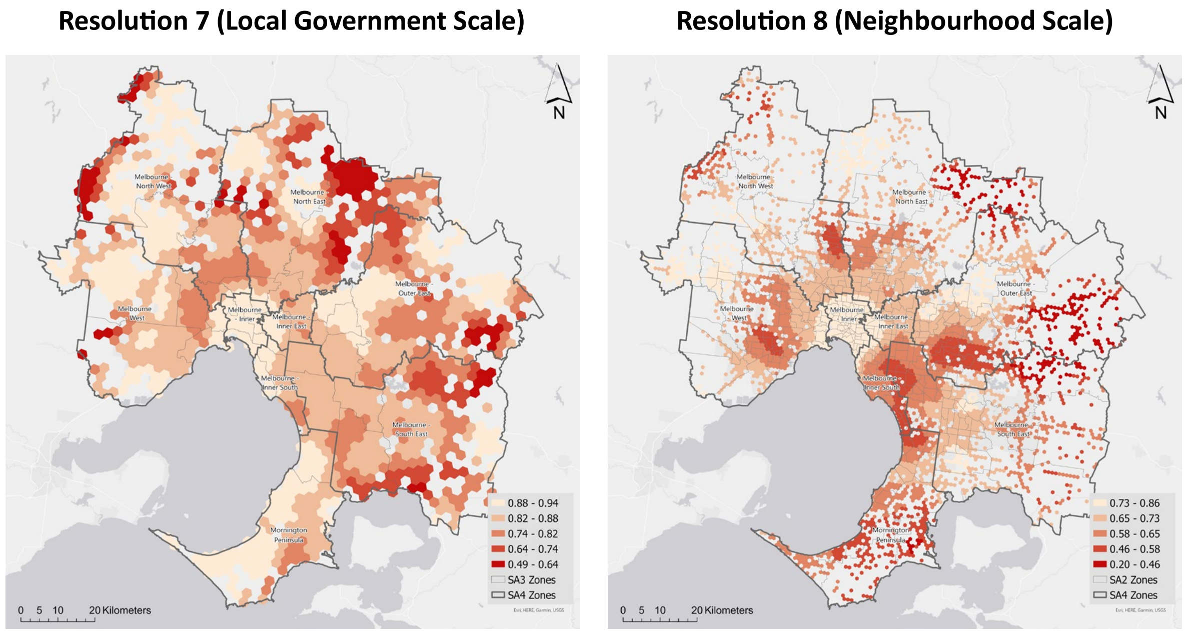
Feature Name	**Coefficient**	**p-value**
Avg. Roads at Intersections	0.301	≈ 0.000
Low Income Population (%)	0.061	≈ 0.000
Avg. Land Surface Temperature	0.046	≈ 0.000
Total Tunnels	0.028	≈ 0.000
Public Transit Usage (%)	0.022	≈ 0.000
Total Bridges	0.008	0.002
Sum Shop POIs	0.006	≈ 0.000
Avg. Speed Limit	0.006	≈ 0.000
Population Aged 65+ (%)	0.001	≈ 0.000
Manager, admin, and prof workers (%)	0.001	≈ 0.000
Link Road Length	≈ 0.000	≈ 0.000
Residential Road Length	≈ 0.000	≈ 0.000
Sum Amenity POIs	≈ -0.000	0.259
Total Roundabouts	≈ -0.000	0.070
Avg. Road Curvature	-0.003	≈ 0.000
Avg. Slope of Roads	-0.005	0.597
Private Transit Usage (%)	-0.006	≈ 0.000
Indigenous Population (%)	-0.013	≈ 0.000
Avg. Built-Up Index	-2.755	≈ 0.000
Avg. Vegetation Index	-3.187	≈ 0.000

Table 5: Resolution 7 (Local Government Scale) GLR Results

R2 = 0.82

Feature Name	**Coefficient**	**p-value**
Avg. Land Surface Temperature	0.174	≈ 0.000
Childless Population (%)	0.017	≈ 0.000
Sum Leisure POIs	0.009	≈ 0.000
English Speaking Population (%)	0.003	≈ 0.000
Male Population (%)	≈ 0.000	0.053
Total Nodes	≈ 0.000	≈ 0.000
Other Road Length	≈ 0.000	≈ 0.000
Total Bridges	-0.002	0.024
Total Tunnels	-0.004	0.593
Sum Tourism POIs	-0.008	≈ 0.000
Avg. Vegetation Index	-3.393	≈ 0.000
Avg. Built-Up Index	-5.322	≈ 0.000

4.5 GWR Local Performance



Resolution 7 (Local Government Scale)

Resolution 8 (Neighbourhood Scale)

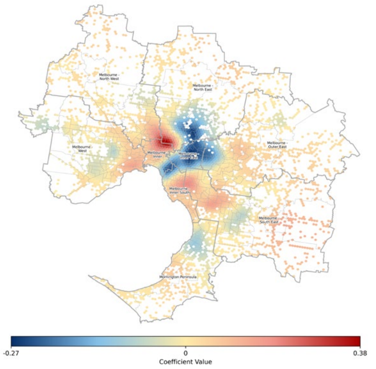
Figure 6: Local R^2 Values of Study Area

Table 6: Summary of Local R^2 Values

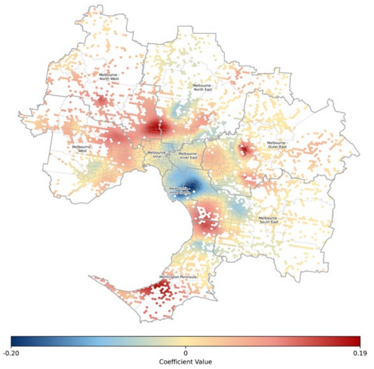
Metric	**Resolution 8 R^2**	**Resolution 7 R^2**
Minimum	0.20	0.49
Maximum	0.87	0.95
Mean	0.65	0.83
Standard deviation	0.10	0.08

1.4 GWR Local Feature Coefficients

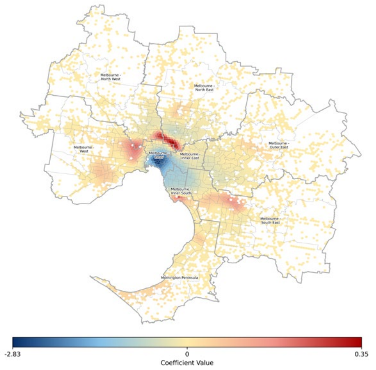
Residential Road Length



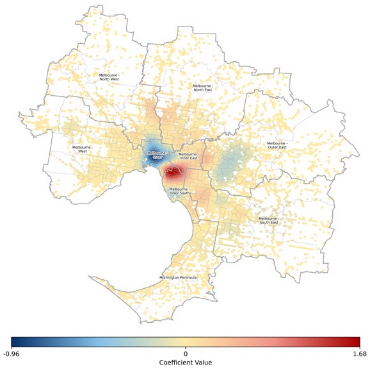
Total Bridges



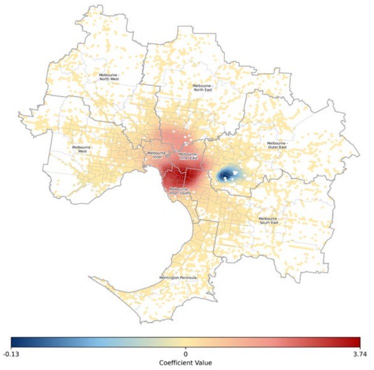
Avg. Road Curvature



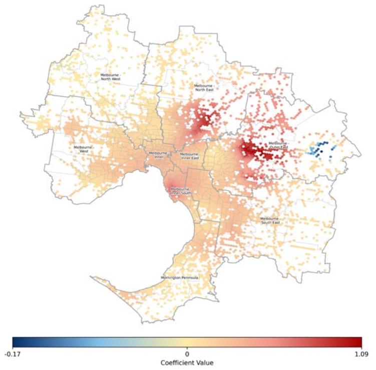
Indigenous Population (%)



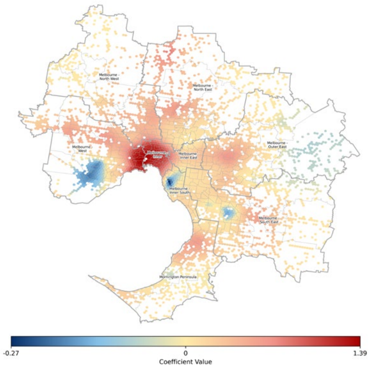
Avg. Roads at Intersections



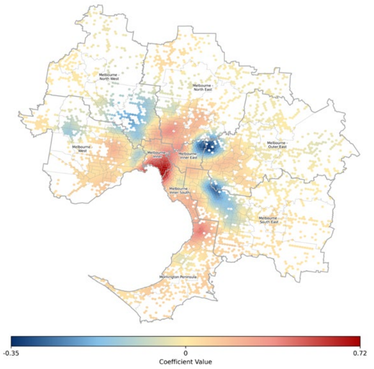
Link Road Length



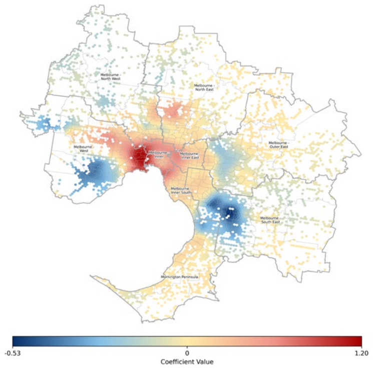
Low Income Population (%)



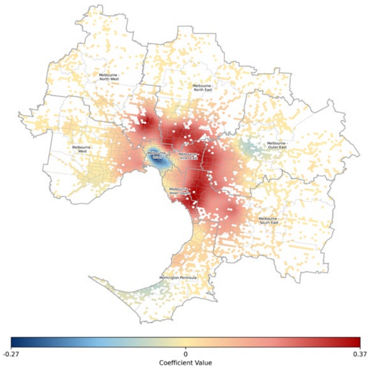
Avg. Land Surface Temperature



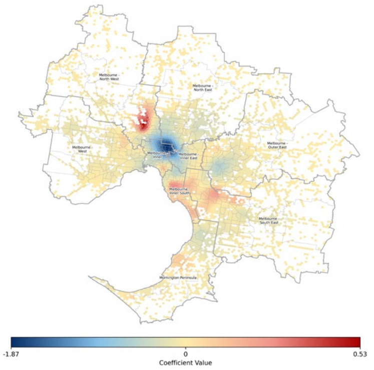
Manager, admin, and prof workers (%)



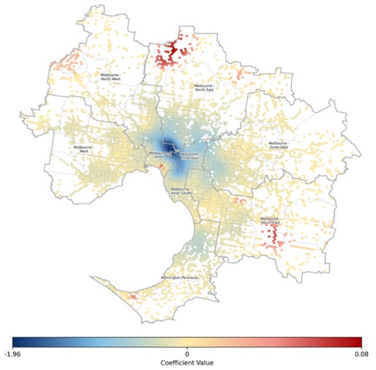
Avg. Speed Limit



Avg. Built-Up Index



Avg. Vegetation Index



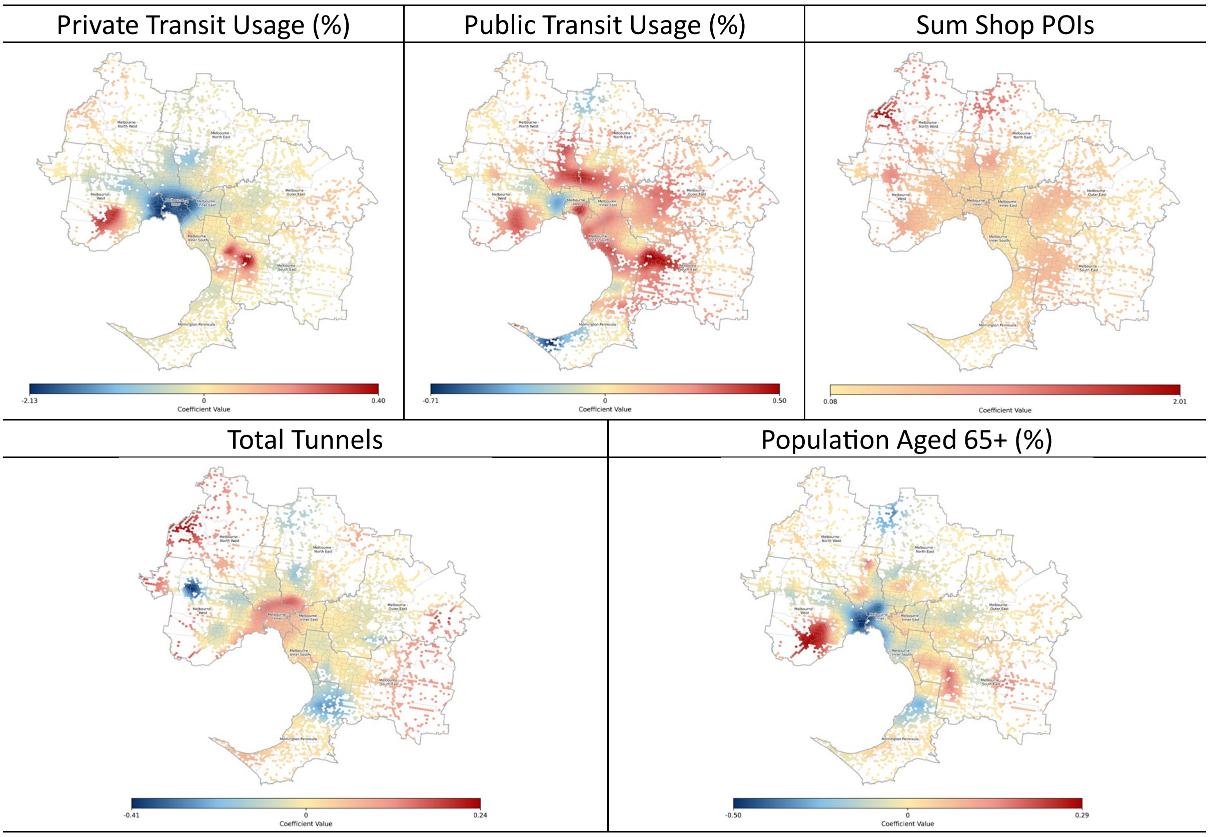


Figure 7: GWR Results for Resolution 8 (Neighbourhood Scale)

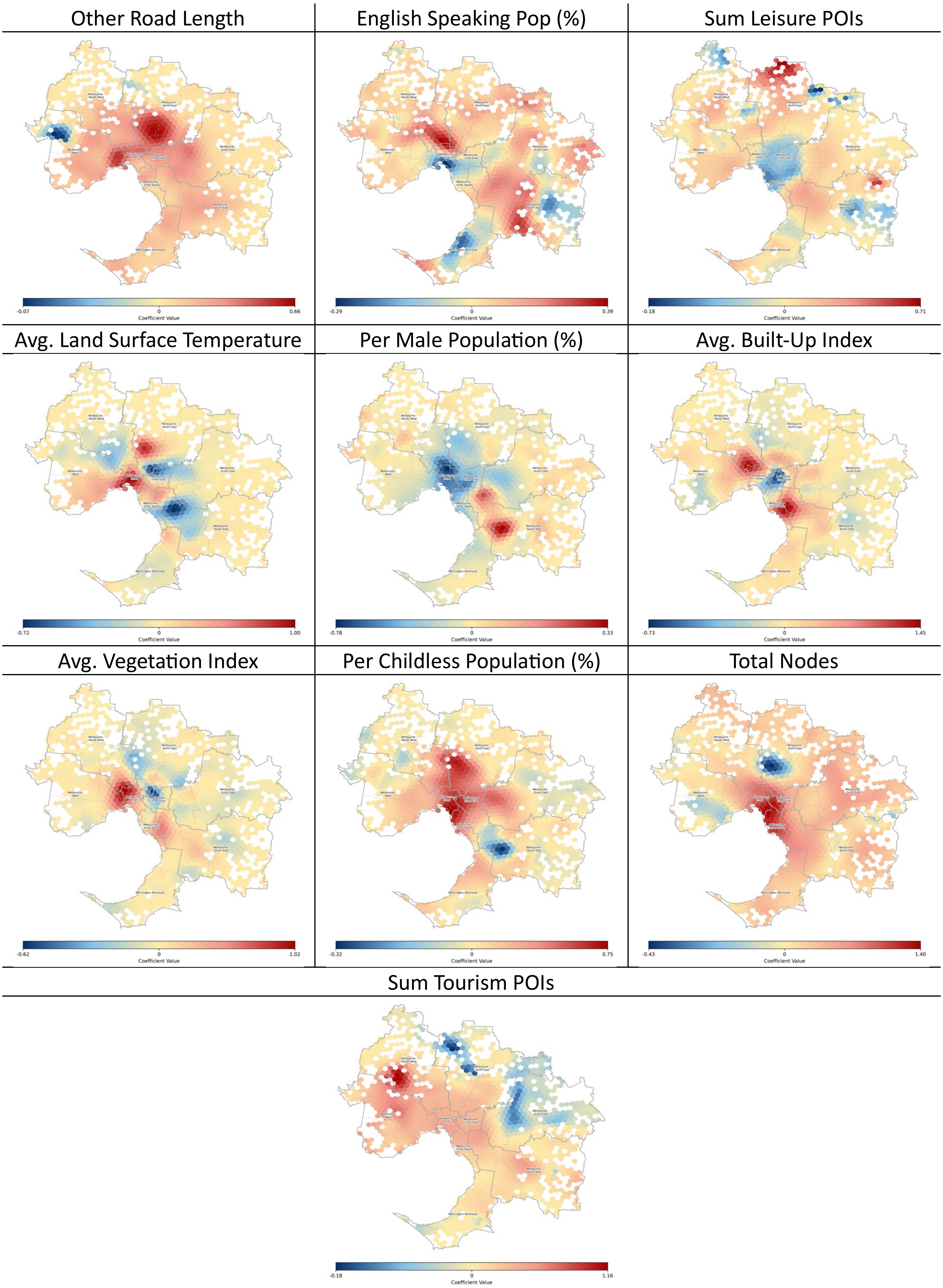


Figure 8: GWR Results for Resolution 7 (Local Government Scale)

5. Discussions and Conclusion

5.1 Forest Prediction Performance

The performance of any machine learning model indicates how confident stakeholders can be in the model’s outputs. Two main metrics that will be discussed include the R2 and Mean Absolute Error (MAE). The R2 shows how much of the variation in the data our model can explain. MAE tells us the average size of the errors our model makes. It gives a tangible sense of the actual difference between the predicted and real values. A high R2 and a lower MAE indicate that a model understands the underlying patterns and provides insight into the quality of the relationships captured. In assessing the performance of the Random Forest model across two scales, clear differences in prediction accuracy and R2 are evident.

At the local government scale an R2 of 0.88 suggests that the model explains 88% of the variance in crash density at this resolution, indicating a relatively strong fit to the data. An MAE of 9.95 indicates that on average, the model's predictions are approximately 10 crashes off from the true total values. With an average of 32 crashes per hexagon throughout the dataset, this performance is strong.

At the neighborhood scale, an R2 of 0.75 suggests the model explains 75% of the variance in crash density. While this is slightly lower than the R2 at Resolution 7, it's still indicative of a solid fit to the data. An MAE of 3.75 indicates that the model's predictions, on average, deviate by roughly 4 crashes from the actual numbers. With an average total crash value of 9.2 per hexagon, the relative performance is commendable.

Given the performance metrics, it's clear that the Random Forest model does not struggle to capture the relationships of total crashes. Stakeholders can have increased confidence in the model's ability to understand and predict crash densities based on these results.

5.2 Key Model-Wide Factors Influencing Crashes

The analysis of feature importance provides valuable insights into the relative significance of different factors affecting the model's decision-making process. For stakeholders, understanding these importance’s can guide decision-making, resource allocation, and strategic planning. Analysing such results from the Random Forest model provides model-wide (Greater Melbourne) insight into important features without the consideration of spatial factors.

Both at the local government and neighborhood scales, the sum of amenity and shop POIs emerges as top-ranking features. Their prominence at both resolutions underscores the universal significance of amenities and shops in urban settings.

Features like 'Other Road Length', 'Residential Road Length', and 'Link Road Length' hold considerable weight, especially at the local government scale. Their importance suggests that road infrastructure plays a pivotal role in shaping urban dynamics.

Demographic features, including 'Population Aged 65+ (%)', 'Childless Population (%)', and 'English Speaking Population (%)', along with socio-economic indicators such as 'Private Transit Usage (%)', exhibit varied importance across resolutions. This variation underscores the need for adaptive policies tailored to the unique demographics and socio-economic realities of each scale.

Features like 'Avg. Vegetation Index', 'Avg. Land Surface Temperature', and 'Avg. Built-Up Index' have moderate importance at the neighborhood scale. Their significance suggests that localized environmental and morphological factors play a role in shaping neighborhood characteristics.

The SHAP summary plots segment the data into categories, encompassing environmental factors, road geometry, demographics, among others. Through these plots, we delved into the influence of each feature on the model's predictions. Within the environmental category, both the density of shop POIs and amenity POIs exhibited a clear positive correlation with crash occurrences. As such, a denser concentration of shops or amenities can indicate a heightened likelihood of crashes. Intriguingly, while both high and low percentages of the childless population were tied to decreased crash incidents, it was predominantly the higher percentages that showed an uptick in crashes. A unique advantage of SHAP values is their ability to reveal the directional impact of a feature on the model—whether it augments or diminishes crash counts. For instance, as the percentage of individuals in managerial, administrative, and professional roles increases, we observed a corresponding decrease in total crashes. Likewise, an increased percentage of private transit usage consistently correlates negatively with crash occurrences across different scales. The SHAP results derived from the two model-wide random forests offer insights into over-arching trends and relationships.

5.3 Local Relationship Performance

While the random forest evaluation metrics provided a model-wide or global view on performance, the GWR model aims to find the spatial variation in such relationships.

At the neighbourhood scale (Resolution 8), the model's R2 values range from 0.20 to 0.87, with an average of 0.65. At the broader local government scale (Resolution 7), the R2 values vary between 0.49 and 0.95, with an average score of 0.83. These R2 values reflect how effectively our data predicts changes in total crashes across different regions. Areas with lower R2 values suggest that there might be other influential factors not captured in our dataset. Generally speaking, a hexagon with a higher R2 can be considered more reliable in terms of the captured relationships.

Both the neighbourhood and local government scale GWR models perform well within the Inner-Melbourne region. This successful capture could be attributed to the higher data density in urban areas, leading to more robust modelling and improved predictions. Both models appear to struggle more around the outskirts of Greater Melbourne though not entirely. This seems to correlate with areas that exhibit patchier distribution of hexagons. This lack of data understandably presents a challenge to the model. At the local government scale, Mornington Peninsula seems to be better represented compared to the neighbourhood scale. In general, there's a noticeable pattern of a decline in performance as crash density diminishes, with certain regions being exceptions. However, in both models, the overall relationship with total road crash occurrences is captured effectively, excluding a few specific, hexagons or regions.

5.4 In Depth Global and Local Discussion

Examining the GLR coefficient results, which indicate both the magnitude and direction of influence on crash occurrences, provides us with insights into the relative importance of different variables within the linear modelling framework. This linear perspective is particularly beneficial when combined with the non-linear, ensemble-based insights from the Random Forest's SHAP values. By comparing the outcomes from SHAP (from Random Forest) and the coefficients from GLR, we can identify consistencies and discrepancies in feature importance and influence across different modelling methods. Such a comparison allows us to understand if certain features consistently emerge as significant influencers regardless of the model type. Additionally, we can recognize if there are non-linear interactions captured by the Random Forest that may not be as evident in the linear GLR model. It's also valuable to determine if the GLR model's assumptions or constraints influence the importance of a feature differently than in a more flexible model like Random Forest. Ultimately, this holistic approach offers a more robust and comprehensive understanding of the underlying relationships in our data, ensuring that our interpretations and subsequent decisions are well-founded.

At the neighbourhood scale, our GLR model identified the average number of roads converging at intersections as having the most pronounced positive correlation. Similarly, our Random Forest model underscored this feature's significance within the area of road geometry factors for this scale. The consistency across both models strengthens the conclusion, especially when the trend isn't easily discernible in the SHAP plots. These insights indicate that as intersection complexity rises, there's a notable uptick in crash occurrences when compared to other road geometry factors. It's advisable for stakeholders to prioritise and address these specific roadway characteristics with added care. An analysis of the spatial distribution reveals a significant positive correlation between average roads connecting to intersections in Melbourne's Inner East, as well as specific sections of both Melbourne's Inner South and South East regions. These areas might possess distinct road designs that potentially contribute to higher crash rates compared to other locations.

The link road length, especially associated with ramps, displays a relatively modest positive association with crash occurrences in our GLR analysis. Yet, this contrasts with the importance it receives in our Random Forest model, suggesting there might be complex non-linear dynamics that a linear framework might overlook. An intriguing hypothesis is the result of competing influences. For instance, certain scenarios within our dataset might exhibit a direct relationship between longer link road lengths and increased crash counts, while others show an inverse association. When these contrasting trends balance each other out, it could result in the GLR model reflecting a more subdued coefficient. This theory is supported when examining our SHAP plots: given the roughly even distribution of blue (representing lower feature values) and red dots (indicating higher values) in the positive contribution. This scenario underscores the value of adopting a diverse modelling strategy to understand the complex nature of the data. In conclusion, the evidence suggests that the total length of road links plays a role in escalating road crashes, potentially becoming more significant beyond a certain threshold of increase. This particular threshold could be further investigated. Upon analysing the spatial results for link road length, a moderate positive correlation is evident in central Melbourne – Inner South. Even stronger positive correlations can be observed in the outskirts of Greater Melbourne. This pattern might indicate that ramps linking to urban roads result in more accidents than those situated in around Inner Melbourne.

Beginning with environmental factors, it's evident that the predominant contributors to crash occurrences include the average land surface temperature at both the neighbourhood and local government levels. Interestingly, the influence is more pronounced at the local government scale. In contrast, our SHAP plots depict a different picture where the average land surface temperature exerts a more substantial impact at the neighbourhood level. This disparity might arise from the Random Forest model's ability to capture complex non-linear relationships that the GLR might overlook at certain scales.

The average vegetation index displays a contrasting trend, exhibiting a pronounced negative correlation across both scales. This observation aligns with the results from the Random Forest SHAP analysis, which emphasises the significance of this factor in relation to other environmental determinants and confirms its negative correlation. Thus, we can deduce that areas with more abundant vegetation generally report fewer crash incidents. A plausible explanation might be the potential inverse relationship between the average vegetation index and the average built-up index, suggesting that areas with higher vegetation might have a lower urban density. However, the relationship between the average built-up index and crash occurrences is somewhat ambiguous. While the SHAP plots do not clearly define the direction of the relationship, the GLR attributes a notable negative correlation to the average built-up index. This seems to challenge the presumption of an inverse relationship between the vegetation index and the built-up index. It's plausible that other underlying factors might be influencing these observed patterns.

Regarding the spatial distribution influenced by land surface temperature, we observe its most pronounced positive influence in certain sectors of the Melbourne – Inner regions. Specifically, the more northern sections within this region exhibit a neutral correlation. Conversely, Melbourne – South East, along with parts of Inner East and Outer East, display distinct clusters with a strong negative influence. Turning to the vegetation index, pronounced negative correlations are evident in the Melbourne Inner East and Melbourne Inner regions. This finding is intriguing and somewhat unexpected given the context and known spatial patterns of the study area. While the precise reasons for this strong negative cluster remain uncertain, a few potential explanations or contributing factors could be considered. These might include localised variations in land use, road infrastructure, or perhaps even data anomalies. Further research or a more detailed examination of local conditions in this specific area might be warranted to gain a clearer understanding of this observed pattern. In summary, these patterns underscore that the associations between crashes and factors like vegetation and land surface temperature are not uniformly distributed but vary significantly across different spatial regions. Certain areas, therefore, might be more vulnerable to these factors than others.

The density of shop POIs exhibited a moderate positive influence at the neighbourhood scale according to the GLR model. In contrast, SHAP plots highlight the random forest model identifying this feature as of paramount importance. Observing the distribution of points, patterns emerge that echo our earlier discussion concerning the road link length feature, suggesting potential competing influences. Employing multiple models for cross-referencing allows us to pinpoint such relationships. Notably, the spatial distribution of the influence of shop POIs throughout the Greater Melbourne region remains fairly consistent. Nearly every area exhibits a subtle positive correlation with increased crashes, underscoring the overarching significance of shopping areas. Consequently, stakeholders are advised to consider policy interventions universally applicable across all shopping regions.

At the local government level, the proportion of the population without children seems to exert a strong influence when compared to other socio-economic factors. This observation is backed by our SHAP analysis, which not only underscores a potent positive correlation but also ranks it as the most influential feature. Consequently, a rise in the childless population is anticipated to correlate with increased crash incidents. Such trends might stem from factors like age and driving experience, potentially hinting at a larger proportion of younger drivers. Additionally, lifestyle differences between those with and without children could play a role: for instance, childless individuals might be on the move more frequently or take more risks, whereas parents might adopt more cautious driving habits.

Spatial analyses reveal pronounced positive correlations in Melbourne - Inner, Inner East, and select areas of Melbourne - Northeast and Inner South. In contrast, other regions display minimal to weak influences, suggesting that other factors might be driving the trends in these areas. The results highlight spatial variations in demographic factors. Certain regions may necessitate a heightened focus on safety measures.

Public transit usage appeared to have a greater influence on increased crashes than private transit usage which had a negative correlation. The SHAP plots support this finding, ranking public transit usage as one of the most important factors. A positive correlation for public transit usage may seem counter intuitive at first however there are some notable reasons. For an example increased activity near transit stops. Buses and trams, for example, make frequent stops and can have different driving patterns than other vehicles. Cars might not always anticipate these patterns, leading to potential crashes. Places with higher public transit usage might also have higher population densities.

The influence of public transit usage on crash occurrences is positive and notably stronger compared to private transit, which is inversely correlated with crashes. This observation is confirmed by the SHAP plots, which identify public transit usage as a key influence. At first glance, the positive association between public transit usage and crash incidents may seem counter intuitive, but several factors offer a clearer perspective. For instance, areas with high public transit activity typically see busy transit stops, especially during peak times. Buses and trams have distinct driving behaviours, frequently stopping to pick up or drop off passengers. Such patterns can be unexpected for other vehicles on the road, elevating the risk of accidents. Moreover, areas with a prominent public transit presence often correlate with higher population densities. This can intensify the vehicular and pedestrian traffic, adding to the complexity of the traffic environment and potentially increasing the likelihood of accidents. Spatial analysis reveals a strong positive correlation throughout Greater Melbourne's inner regions. A majority of the area exhibits at least a moderate positive correlation, underscoring its importance. Particularly significant clusters include regions in Melbourne – South East and Melbourne – Inner. Stakeholders are advised to prioritize these areas in the context of public transit planning.

Our findings align with existing literature in several key areas. Firstly, our identification of intersection complexity as a significant predictor of crash occurrences corroborates studies emphasizing the role of intersection design in road safety (Mussone et al., 2017). This supports the notion that high-complexity intersections may introduce increased risk due to factors like higher traffic volume and complex manoeuvring requirements. Our findings concerning environmental factors, specifically the association between temperature and crash occurrences, resonate with studies exploring the impact of weather conditions on road safety (El-Basyuni et al., 2014, Gao, J. *et al*). The significance of vegetation index in reducing crash incidents aligns with research highlighting the safety benefits of green infrastructure and vegetation in urban areas (Murphy and Xia, 2016). Regarding socio-economic factors, our identification of the proportion of the population without children as a strong influencer of crash occurrences corresponds with studies exploring the relationship between demographic factors and road safety outcomes (Ellison et al., 2015, Hasanat-E-Rabbi et al., 2021). Overall, our findings contribute to the existing body of knowledge on road safety by providing empirical evidence of the specific factors influencing crash occurrences in Greater Melbourne. Our results underscore the importance of considering both linear and non-linear modelling techniques to gain a comprehensive understanding of the underlying relationships in road safety data. This approach aligns with recent trends in road safety research and emphasizes the need for a multifaceted approach to inform effective road safety interventions.

The profound impact of vehicular accidents on public health, economic stability, and societal well-being underscores the critical need for comprehensive and data-driven preventive strategies. This study embarked on an ambitious journey to explore the multitude of factors contributing to road accidents in Greater Melbourne using advanced machine learning and geospatial analytics.

Our use of a multi-model approach, integrating RF with SHAP, GLR, and GWR, in conjunction with hexagonalised datasets, has unearthed nuanced patterns and delineated high-risk zones that conventional statistical methods may overlook. Such discoveries accentuate the transformative potential of contemporary analytical techniques in amplifying our grasp of the dynamics of road safety. This enriched understanding, stemming from our multi-model methodology, has proven pivotal in unmasking often overlooked complexities.

Greater Melbourne, with its diverse socio-economic and environmental landscape, presents unique challenges. However, our research illuminates the potential for urban planners, policymakers, and stakeholders to harness data in crafting effective preventive measures. It was found that complex intersections, particularly in specific Melbourne regions, have a notable association with crash rates. Revisiting problematic areas or incorporating better safety planning for new intersections could be beneficial. Regions with a higher childless population see more crashes, suggesting a potential focus for safety campaigns. Additionally, areas with increased public transit activity, especially in Melbourne's Inner and Southeast regions, require heightened safety measures potentially around transit stops.

While this study offers significant insights, there remains a vast scope for future research. Exploring more specific contributing factors could further refine our understanding. Additionally, a comparative study with other urban regions could provide broader perspectives on urban road safety. Furthermore, investigation of different hexagon scales could be beneficial. This study had several limitations. First, enhanced domain knowledge in crash prevention might have better guided the assignment of features to their respective scales. The most significant limitation, however, stems from our reliance on open-source data from OSM, which could introduce inaccuracies into our models. Further refinements in model selection could be achieved with expertise, providing a deeper understanding of the advantages and disadvantages of each model.

In an era where data-driven decision-making is paramount, our research underscores the necessity of merging technology with urban planning. Only through such integrative efforts can we hope to make our urban environments safer and more resilient to the challenges of modern-day transportation.

References

Al-Mistarehi BW, Alomari AH, Imam R and Mashaqba M (2022) 'Using Machine Learning Models to Forecast Severity Level of Traffic Crashes by R Studio and ArcGIS', Frontiers in Built Environment, 8, doi:10.3389/fbuil.2022.860805. https://www.frontiersin.org/articles/10.3389/fbuil.2022.860805

Álvaro B-R, Francisco M-R and Francisco M (2019) 'Investigation of the consequences of the modifiable areal unit problem in macroscopic traffic safety analysis: A case study accounting for scale and zoning', Accident Analysis & Prevention, 132:105276, doi: https://doi.org/10.1016/j.aap.2019.105276. https://www.sciencedirect.com/science/article/pii/S0001457519304385

Asadi M, Ulak MB, Geurs KT, Weijermars W and Schepers P (2022) 'A comprehensive analysis of the relationships between the built environment and traffic safety in the Dutch urban areas', Accident Analysis & Prevention, 172:106683, doi: https://doi.org/10.1016/j.aap.2022.106683. https://www.sciencedirect.com/science/article/pii/S0001457522001191

Dumbaugh E and Zhang Y (2013) 'The Relationship between Community Design and Crashes Involving Older Drivers and Pedestrians', Journal of Planning Education and Research, 33(1):83-95, doi:10.1177/0739456X12468771, accessed 2023/10/10. https://doi.org/10.1177/0739456X12468771

Elyassami S, Hamid Y and Habuza T 2020, Road Crashes Analysis and Prediction using Gradient Boosted and Random Forest Trees, 5-12 June 2021, 2327-1884.

Ellison, A. B., Greaves, S. P., & Bliemer, M. C. (2015). Driver behaviour profiles for road safety analysis. Accident Analysis & Prevention, 76, 118-132.

El-Basyouny, K., Barua, S., & Islam, M. T. (2014). Investigation of time and weather effects on crash types using full Bayesian multivariate Poisson lognormal models. Accident Analysis & Prevention, 73, 91-99.

Gao, J. et al. The association between meteorological factors and road traffic injuries: A case analysis from Shantou city China. Sci. Rep. 6(1), 1–10 (2016)

Hasanat-E-Rabbi, S., Hamim, O. F., Debnath, M., Hoque, M. S., McIlroy, R. C., Plant, K. L., & Stanton, N. A. (2021). Exploring the relationships between demographics, road safety attitudes, and self-reported pedestrian behaviours in Bangladesh. Sustainability, 13(19), 10640.

Lee D, Guldmann J-M and von Rabenau B (2018) 'Interactions between the built and socio-economic environment and driver demographics: spatial econometric models of car crashes in the Columbus Metropolitan Area', International Journal of Urban Sciences, 22(1):17-37, doi:10.1080/12265934.2017.1369452. https://doi.org/10.1080/12265934.2017.1369452

Liu X and Xia J (2015) 'Locally analysing the risk factors for fatal single vehicle crashes hot spots in Western Australia', International Journal of Crashworthiness, 20(6):524-534, doi:10.1080/13588265.2015.1055649. https://doi.org/10.1080/13588265.2015.1055649

Mussone, L., Bassani, M., & Masci, P. (2017). Analysis of factors affecting the severity of crashes in urban road intersections. Accident Analysis & Prevention, 103, 112-122.

Murphy, A., & Xia, J. (2016). Risk analysis of animal–vehicle crashes: a hierarchical Bayesian approach to spatial modelling. International Journal of Crashworthiness, 21(6), 614-626.

Pour AT, Moridpour S, Rajabifard A and Tay R (2017) 'Spatial and temporal distribution of pedestrian crashes in Melbourne metropolitan area', Road & Transport Research, 26(1):4-20, accessed 2023/10/09. https://search-informit-org.ezproxy.lib.rmit.edu.au/doi/10.3316/informit.820985790367486

Rhee K-A, Kim J-K, Lee Y-i and Ulfarsson GF (2016) 'Spatial regression analysis of traffic crashes in Seoul', Accident Analysis & Prevention, 91:190-199, doi:https://doi.org/10.1016/j.aap.2016.02.023. https://www.sciencedirect.com/science/article/pii/S0001457516300562

Tavris DR, Kuhn EM and Layde PM (2001) 'Age and gender patterns in motor vehicle crash injuries: importance of type of crash and occupant role', Accident Analysis & Prevention, 33(2):167-172, doi:https://doi.org/10.1016/S0001-4575(00)00027-0. https://www.sciencedirect.com/science/article/pii/S0001457500000270

Wang Y and Zhang W (2017) 'Analysis of Roadway and Environmental Factors Affecting Traffic Crash Severities', Transportation Research Procedia, 25:2119-2125, doi: https://doi.org/10.1016/j.trpro.2017.05.407. https://www.sciencedirect.com/science/article/pii/S2352146517307147

World Health Organization. (2022). Road traffic injuries. Available at: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries (Accessed 2023/10/10).