Road Safety Data Analysis

Kathlyn Anne Ycong

School of Computer Science / Department of Statistics

The University of Auckland

Supervisor: Simon Urbanek

A dissertation submitted in partial fulfillment of the requirements for the degree
of Master of Professional Studies in Data Science,
The University of Auckland, 2022

Abstract

Road safety is a critical concern for both government and businesses worldwide compelling them to act on reducing traffic accidents. Analysing road safety enables organisations to identify factors associated with traffic accidents to develop preventive measures for example, implementing policies and improving road systems. This study aims to determine the necessary road safety measures, data, and geospatial analysis that will be bundled together in an R package that is accessible and capable of reproducing road safety analysis for a given dataset. We have used open-source data extracted, 2018 main means of work from Statistics New Zealand and crash data from Waka Kotahi New Zealand Transport Agency, which has facilitated us to simulate each commuter's journey and analyse the road safety risk. We have determined that there is an association between the distance travelled by the commuter and the likelihood of being exposed to a traffic accident. However, considerations should be made on the type of road and section of the road where crashes occurred and the changes in the working environment have a relative effect on road safety.

Acknowledgements

Firstly, I would like to thank my supervisor, Professor Simon Urbanek, for his unwavering support and guidance, for imparting his wealth of knowledge in Data Science and spatial analysis, and for ensuring that I have access to tools and resources to complete this research. I am grateful that his high standards have made me better at what I do.

I would also thank Dr Shrividya Ravi and the Ministry of Transport for offering me this opportunity to work on this project and for guiding me through the project.

Finally, I thank my husband and my family for their support.

Table of Contents

Chapter 1
Introduction

According to the World Health Organisation (WHO), the global death toll from traffic accidents is 1.35 million annually, causing social and financial implications [2]. In New Zealand, at least one person dies, and seven others sustain serious injuries from traffic accidents every day [1]. This makes it crucial for businesses and governments to understand and take action toward road safety, a critical concern worldwide [2]. To do so, different methodologies can be applied to navigate the problem of road crashes and inform policy with solutions. Data and analysis are essential in understanding road crashes. It can enable an in-depth understanding of potential factors contributing to the problem and ultimately support prevention measures such as developing and implementing policies and changes to road systems that mitigate road safety risks [3].

R packages are valuable tools for sharing work with others by bundling code, data, documentation, and tests [4]. The data and code in an R package can be used to analyse potential factors to road safety problems. The benefit of packages is that a wide variety of available packages can be used for data and analysis. Bundling work into a package allows others to reproduce work and contribute by iterating and evolving the research [4]. For example, *ghroute*, *spdep* and *sf* are some of the R packages that are useful in analysing data containing geospatial objects, which is fundamental in analysing road safety.

Trafficalmr is an R package developed by a team of researchers from the University of Leeds to automate the access and analysis of road safety data that will support the Department for Transport in the United Kingdom to implement traffic calming measures that reduce the casualty rate from road traffic accidents [5]. However, this package is restricted to the United Kingdom road safety context and thus cannot be applied to New Zealand road safety analysis. This research aims to develop an R package in which the data used in analysing road safety in New Zealand is accessible as well as the methods used to analyse road safety are also accessible through R functions irrespective of the country or context, such that it is all bundled in an R package. To develop this R package, this research uses the actual journey to work data from Statistics New Zealand and crash data from Waka Kotahi New Zealand Transport Agency. Moreover, this research intends to contribute by analysing New Zealand road safety, gathering insights, and posing future directions.

1.1 Overview of Approach

This research focuses on automating road safety analysis. The main idea is to leverage available data and existing R packages to conduct data wrangle and calculate various road safety measures. To achieve this analysis, there are a series of steps required. Firstly, this research determines the datasets needed for analysing road safety in New Zealand. A wide variety of open-source data provided by Statistics New Zealand, New Zealand Transport Agency and Auckland Transport are easily accessible. The correct data are crucial in road safety analysis [3]. Second, various road safety measures for calculating road safety risk are established. These measures are exposure, road safety risk and journey to work risk. These road safety measures are then translated into R functions to automate and reproduce the analysis. This involves simulating commuters' journey from their usual residence to the workplace. Third, the performance of the simulated route is evaluated, which in effect validates the road safety risk. Fourth, the gathered insight using the R packages is summarised and visualised through tools such as *mapview*, *leaflet* and *tmap*. Lastly, the road safety analysis R functions and data are packaged in an R package to reproduce the analysis and be easily shared.

1.2 Contributions

This research makes the following contributions to road safety analysis and application:

- Collate the datasets required to analyse road safety in New Zealand. Data analysis is performed for these datasets.

- Introduce approaches to calculating road safety risk and exposure.

- Evaluating the performance of the simulated routes with the data from the Auckland Transport traffic counter.

- Summarise and visualise insights from the road safety analysis and spatial autocorrelation.

- Develop an R package containing the dataset and function to automate road safety analysis applicable to a broader context.

1.3 Scope and Structure

This dissertation proposes developing an R package to automate access to data and analysis for road safety. We also propose using the distance travelled as road safety exposure to calculate road safety risk. Following this introductory chapter, the rest of the work is organised as follows.

Chapter 2 focuses on reviewing the literature on road safety, the policies implemented to address road safety, various road safety measures, and related work on bundling road safety analysis and spatial autocorrelation.

Chapter 3 focuses on analysing the datasets used for analysis. The chapter mainly focuses on exploring the different data useful in analysing road safety.

Chapter 4 consists of research methodology and assumptions in analysing road safety. Other information includes validating the methods used in this research and geospatial data analysis.

Chapter 5 discusses the results from the data analysis that will be conducted and validates the results using the traffic counter from Auckland Transport. The results from spatial autocorrelation of local clusters are also discussed in this chapter.

Chapter 6 provides an overview of the '*motroadsafety*' R package. This chapter delves into the various datasets and R functions bundled in the package. It also shows how to access the R package from GitHub, the data, how the R functions work, and the parameters required for each R functions in the package.

Chapter 7 concludes this dissertation, discusses the limitations and present some directions for future work.

Chapter 2
Literature Review

2.1 Road Safety

Roads are vital in connecting people and providing access to education, work, and recreation. Roads allow people and products to move around and communities to interact, thus road systems must deliver their purpose [1]. However, roads can be perilous. World Health Organisation (WHO) reported that road traffic accidents are the 8th leading cause of death for people of all ages with 1.35 million deaths reported annually [2]. In New Zealand, people spend on average an hour a day travelling with one person killed every day and seven others seriously injured due to road crashes [1]. Road traffic crashes also have social and financial implications such as medical costs, production losses, property damage, settlement costs and many more that severely burden a country's economy. It is estimated that these costs range between 1 to 5 % of the nation's gross domestic product (GDP) [3]. Every death or serious injury on our roads is a call to act, investigate, diagnose and address [1]. Thus, road safety should be a critical investment priority and instead of being traded off against other urgencies, should be managed through effective designs and policies.

The United Nations have included road safety as part of the 17 sustainable development goals [3]. This road safety goal involves halving the number of deaths and injuries globally caused by road traffic accidents and providing access to transportation systems that are safe, affordable, and sustainable and caters to the vulnerable such as women, children, disabled and elderly persons [3]. According to Nævestad, Phillips, and Elvebakk [6], road transportation poses a significant risk during a regular working day, especially for employees where driving is their profession. Between 20% and 40% of the work-related incidents are due to road traffic accidents; this increases to 60% if we account for the journey to and from work. Consequently, these sustainable development goals have been adopted by many governments and private organisations [3].

2.2 Road Safety Policies

Road safety policies have contributed to improving human behaviours, designing and planning safer infrastructure, vehicle safety and civil infrastructure [3]. However, implementation of road safety policies requires support from the private sector, society and, more importantly, strong leadership in the public and political sector that can evidently afford good results [3]. For example, the seat belt policy implemented in 1984 has consistently reduced fatal crashes by 15.6%. This demonstrates that policies significantly influence road safety [7]. Several countries have developed innovative and radical policies, such as Sweden's Vision Zero, introduced in 1997, influencing other countries [8]. For example, New Zealand has also implemented a similar Vision Zero policy, the 'Road to Zero'. It was implemented with the same vision that no one is killed or seriously injured from road accidents by establishing road safety on New Zealand roads, streets, footpaths, cycleways, bus lanes and state highways over the next ten years [1]. Although these policies have positive results, a comprehensive assessment of the implementation and outcomes must be executed [8].

Several tools have been created to support the development of road safety policies [9, 10]. Elvik [10] cited that economists have used the cost-benefit analysis and the underlying principle of efficiency to enable policymakers to develop policies to solve road safety problems by trying to minimise costs and resources while simultaneously maximising the benefits [10]. An effective and efficient approach to addressing road safety problems and developing and implementing road safety policies is through a data-driven approach or "going fishing where the fish are" [3]. A data-driven approach entails investigating and analysing the data to understand crash causes, the factors that contributed to the occurrence of the risk, determine the crash severity and identify locations where road improvements are necessary to reduce collisions. Wegman [3] highlighted that the traditional approach to road safety for countries with mature road safety policies has limitations. Firstly, the inherent risk from a combination of the physical vulnerability of the human body and the kinetic energy poses unsafe road conditions. Although transportation systems are designed and built to foster safe conditions, it does not warrant reducing the risk of traffic accidents due to human error. Lastly, traditional policies are subject to the law of diminishing returns, where traditional interventions become less effective and efficient over time [3]. This suggests that road safety measures are invaluable in benchmarking and monitoring the effectiveness and performance of policies and their implementation to address road safety [9].

2.3 Road Safety Measures

Road safety measures are instrumental to achieving road safety targets, substantially reducing road fatalities [11] by improving transportation safety and determining public health priorities [12]. According to Wong and Sze [11], a vital component of a road safety strategy is a quantified target. However, measuring road safety is inherently complex [8]. It is crucial to determine the elements and confounding factors contributing to measuring road safety and the various interventions implemented in a single period [8].

Risk, exposure, and consequence are the three fundamental elements of designing road safety measures [13]. The road safety risk is the probability of a traffic accident occurring [12] as defined in equation 1 [14].

$$\text{Road safety risk} = \frac{\text{consequence}}{\text{Exposure}} \qquad (1)$$

Where consequence is defined as the road safety outcome, typically the total number of fatalities or serious injuries [14], on the other hand, exposure is defined as a road user's likelihood of being involved in dangerous or hazardous situations [15]. Pei, Wong, and Sze [15] have argued that exposure is vital in quantifying road safety risk as it will help in understanding the different factors that have influenced the occurrence of a collision, for example, vehicle speed [15]. Exposure has been expressed differently depending on the data available, such as the traffic volume, conflicts, the amount of distance travelled, amount of time travelled, and the total population or fuel consumed [14, 15].

The most widely used definition of exposure is the distance travelled in kilometres for each type of travel mode, either by vehicle or on foot [12]. Merlin, Guerra and Dumbaugh [16] have cited the relationship between exposure and the likelihood of a collision, such as when a person increases travel either by distance or time. It increases its probability of being involved in a traffic-related injury or death [16]. Moreover, a study by conducted Burdett, Starkey, and Charlton [17] has established the "close to home" effect for both male and female drivers. This suggests an increased likelihood of a crash at a shorter distance because of the driver's familiarity and overconfidence on these roads, where drivers demonstrate unsafe behaviour such as altered visual search and inattentional blindness [17].
On the other hand, a disadvantage of using the amount of time travelled is that exposure on the road might be reduced when speed increases, thus reducing the likelihood of being involved in a crash. A study conducted by Pei et al. [15] modelling the relationship between speed and occurrence of a collision concerning distance and time exposure revealed a positive correlation between speed and crash risk when considering distance as the exposure while negative when considering time as the exposure [15]. Thus, a suitable exposure must be used to make road safety analysis more effective.

Data is crucial in quantifying road safety. It supports the quality of road safety analysis. Technologies such as advanced intelligent transport systems (ITS) and GPS probes have helped gather data for road safety analysis [15]. Road design, weather conditions, and temporal distribution for the association between speed, speed dispersion and traffic safety are other helpful data points in modelling crash risk [15]. However, it is known that specific data have distinctive issues. For example, certain crash types are significantly underreported, a limitation of road safety analysis [3]. Cycle crashes are an example of being a victim of underreporting. Many countries have recognised this issue and have taken an approach to address this [3].

2.4 Geocomputation

Geocomputation has been used to understand and solve ecology, marketing, and transportation problems [18]. Geocomputation involves using geospatial data and tools like R to analyse and solve problems programmatically such as calculating distance travelled and manipulating different geometry types. It was developed from the geographic information systems which emerged in the 1960s. The main advantage of using geocomputation is that it is reproducible and modularised, which means that solutions or analyses developed are transferrable to other contexts [18].

The Coordinate Reference System (CRS) is fundamental in working with geospatial data. It is a framework that relates Earth's surface to the spatial data elements [18]. There are two reference systems in CRS, Geographic coordinate system and Projected coordinate systems. In geographic CRS, the Earth is modelled either spherical or ellipsoidal using angular distance to obtain the geospatial coordinate system. Longitude and latitude are the values used to determine a spatial location on Earth's surface. The longitude represents the East-west direction in angular distance from the Prime Meridian plane. Latitude represents the North or South direction in angular distance from the equatorial plane. Spherical geographic models assume that Earth is a perfect sphere which is inaccurate because the Earth is not a sphere [18]. Hence, the ellipsoidal geographic model is commonly used since it closely represents the shape of the Earth.

The project coordinate reference system represents Earth by converting the geographic CRS three-dimension into a specific surface using the Cartesian coordinates systems using X (Easting) and Y (Northing) as values; this can measure distance in units, for example, meters. It is essential to note that the caveat of transitioning from geographic to projected CRS of Earth's surface distorts area, direction, distance, and shape [18] because there is no mathematical way that a three-dimension physical world can represent a two-dimensional flat surface [19]. However, the projected CRS can preserve either one or two properties as mentioned earlier. For example, equal-area projection preserves area property, azimuthal projection preserves the direction, equidistance preserves the distance and lastly, conformal preserves the local shape [18].

The project coordinate reference systems have three types, conic, cylindrical and planar. Conic projection is suited for maps of mid-latitude areas since distortion is minimised along the tangency lines and rises with distance from those lines because Earth's surface is projected onto a cone along a single line or two lines of tangency. Cylindrical projections are generally used to map the entire world. It maps the surface of Earth onto a cylinder. Like conic, it uses a single line or two lines of tangency to create the projection. Lastly, planar projection is usually used in polar mapping regions. It represents Earth as a flat surface [18].

Each country has a varying projection system due to differing landscapes thus an appropriate CRS must be used for a given spatial data to ensure quality and accuracy in any geographic computation [19]. The datum is a component in CRS that contains information on the relationship between the Cartesian coordinates and location on Earth's surface, considering the variations in the Earth's surface, such as mountain ranges for a given local CRS [18]. Thus, the correct coordinate system must be used according to the location and purpose to ensure data quality.

2.5 Trafficalmr

The SaferActive is a project funded by the United Kingdom Department for Transport to support road safety policies aiming to increase active travel, i.e., cycling and walking, while reducing casualties year on year [5]. The project has revealed that road safety interventions have reduced casualty rates. The project scope included analysing data from various sources, data modelling and developing an R package, called *trafficalmr*, to automate access to data and road safety analysis [5].

The analysis conducted by the SaferActive project included the casualty data containing fatalities from serious to slight casualties from cycling and pedestrian that occurred between 2009 to 2013. They have also used the 2011 census data to estimate the route and the distance travelled, assuming that the commute was one-way on a single day in 2011 that occurred during rush hour (i.e., 7:30 to 9:30 and 16:30 to 18:30) on weekdays. The project used *Cyclestreets.net* fast-routing algorithm to construct the route network. The project also uses traffic calming intervention data extracted from OpenStreetMap (OSM) to measure its effectiveness in increasing road safety. It was interesting to determine the interaction between using open-source datasets to analyse road safety [5]. The Traffic for London (TfL) data which consists of the traffic counts from 2015 to 2019, was also used to validate the estimated cycle count exposure [5].

The Saferactive defines exposure as the estimated amount of travel in kilometres. The road safety outcome is calculated as the total accidents during rush hour on a weekday divided by two to represent a one-way journey. It is divided by 5 to mean a single year and then by 261 to represent 261 working days in a year in the United Kingdom. Road safety risk with the unit killed and seriously injured by billion kilometres travelled (KSI/bkm) is then calculated by dividing the road safety outcome by exposure. The routes, represented as line segments, that intersect with the London boroughs, represented by a spatial polygon, are aggregated by the total number of commuters that have traversed on that route by each borough to compare road safety risk. The project has also noted that corrections were made to the crash data due to changes to the accident reporting system by the police force that resulted in amending the proportion of casualties [5].

The project has revealed a strong negative relationship between km cycled and KSI/bkm. Commuter travelling by cycling tends to cycle in the inner boroughs of London and has a higher total number of casualties in contrast to the London outer boroughs where there are fewer cyclers and higher KSI/bkm. For pedestrians, the analysis reported that 25.9% of the casualties are severe or fatal compared to 19.1% for cyclists. The outermost inner London Boroughs have the highest rates of KSI/bkm. The research also indicated that junctions must be analysed separately from roads as junctions tend to be hotspots for collisions [5]. Furthermore, the increasing trend of working from home has a significant effect on analysing road safety because the pandemic has amplified the proportions of adults working from home, which will likely remain in effect for the long term [5].

2.6 Spatial Autocorrelation

Spatial Autocorrelation is a concept used to describe the spatial variance in exploratory spatial data analysis [20]. Similar to correlation, positive autocorrelation refers to similar spatial units, while negative autocorrelation refers to less similar spatial units for a given variable [18, 19]. It is generally used for exploratory spatial data analysis in ecology [21], and it helps determine spatial clusters, i.e. hot spots and cold spots that are beneficial in analysing geospatial data [20]. The *spdep* package available in Comprehensive R Archive Network (CRAN) has been developed to conduct spatial data analysis using spatial autocorrelation [20].

Spatial autocorrelation has two methods global and local spatial autocorrelation. The global spatial autocorrelation refers to a global variance in the dataset and specifies the degree of clustering between neighbouring spatial units. *Moran's I* is a statistical test on spatial randomness typically used to determine the existence or non-existence of global spatial autocorrelation where a value of -1 indicates perfect dispersion, 0 refers to perfect randomness, and +1 refers to perfect clustering of similar values [20]. It is a linear clustering technique [22]. The Z-score and p-value are used for the statistical hypothesis, where a very small p-value refers to rejecting the null hypothesis. On the other hand, a significant p-value means that we fail to reject the null hypothesis [22]. *Moran's I* coefficient is defined in equation 2 [22].

$$I = \frac{\mathrm{n} \sum_{i=1}^{n} \sum_{j=1,}^{n} \omega_{i,j} \, z_i \, z_j}{S_o \sum_{i=1}^{n} z_i^2} \qquad (2)$$

Where n is the total number of samples, ω_i is the spatial weight between the spatial unit i and neighbouring spatial unit j, z_i is the deviation of a given attribute for feature i from its mean $x_i - x$, and S_o refers to the spatial weight aggregation.

Local spatial autocorrelation is another method in spatial autocorrelation that detects variability and divergence with one particular spatial unit in contrast with global spatial autocorrelation, which looks at the overall trend of all the spatial units. Local indicators of spatial associate (LISA) quantify the extent a spatial and its neighbouring spatial unit are similar or dissimilar, generalising the concept of *Getis-Ord Gi* and *Getis-Ord Gi** statistics on the presence of hotspots [22]. It is also valuable for identifying influential observations and outliers by decomposing global statistics such as *Moran's I* coefficient into their local components that are helpful with data quality [20]. LISA classifies the spatial areas into four groups high values near to high values (HH), low values with the nearby values (LL) and low values with high values in their neighbourhood and vice versa [20].

Spatial weights help determine geometry's neighbourhood. Three types of representation of spatial weights need to be considered. Firstly, the spatial association represents the distance decline function. Spatial nearness is a geometric indicator that represents the contiguous spatial units. Lastly, the descriptive expression of the spatial association is included in the data set. There are various methods for determining spatial weights [20]. For instance, Queen and Rook are examples of methods for determining spatial weights [23].

Chapter 3
Dataset Analysis

3.1 Crash Data

The crash data is sourced from the Waka Kotahi Crash Analysis System (CAS), which records all road traffic crash reported to the New Zealand Police to analyse road safety risk. Various organisations have used CAS data with the broad aim of improving road safety [24]. For example, the New Zealand government uses the CAS data to support the adoption of the 'Road to Zero' strategy. The vision of the said strategy is that no one is killed or seriously injured in road crashes while travelling on New Zealand roads. The data has also facilitated transportation policy, designing, and prioritising road safety improvements and monitoring their effectiveness [25].

This study is interested in crashes that occurred in 2018 in alignment with the 2018 census focusing on the main means to travel to work data. According to Waka Kotahi New Zealand Transport Agency, non-injury crashes are vastly underreported [24]; however, it is used in this study to analyse the risk exposure for each unique journey made by the driver. The projection coordinates in the CAS data are used to determine the specific location of the crash and to analyse if the crash is within close proximity to the journey made by the driver given a particular distance. This research will use the New Zealand Geodetic Datum 2000 (NZGD2000) as the projection reference system or EPSG code 2193, the standard for all geospatial data in New Zealand [26].

Figure 3.1 is a time series plot that shows the trend of the total number of crashes for each crash year by crash severity. Minor crashes and non-injury crashes are the predominant types of crashes reported in CAS data. From Figure 3.1, we can notice that around 2013, there was a dip in crashes and in 2018, there was a peak of crashes reported for all crash severity.

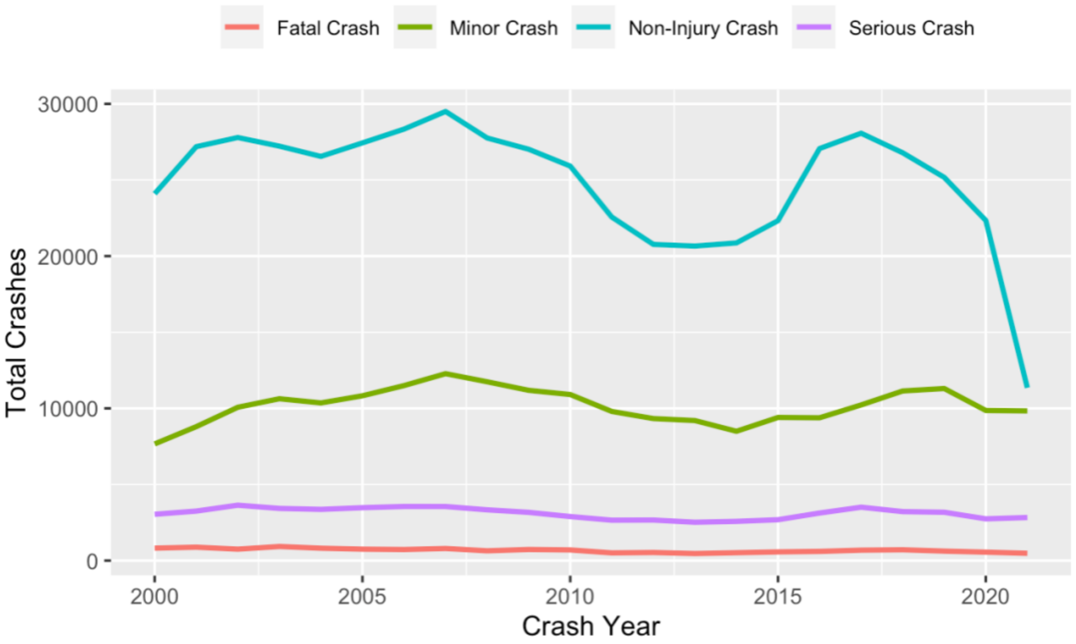


Figure 3.1: Total crashes in New Zealand for every Crash year by Crash Severity

Injury crashes typically occur at road junctions, as seen in figure 3.2. It is apparent that the road junctions have a high density of crashes. This aligns with the insights from the SaferActive project that indicated that junctions are known collision hotspots [5]. It is also noticeable in figure 3.2 that busy roads such as the North-western motorway and Southern motorway have high-density crashes; these are some insights that need to be considered in the road safety analysis.

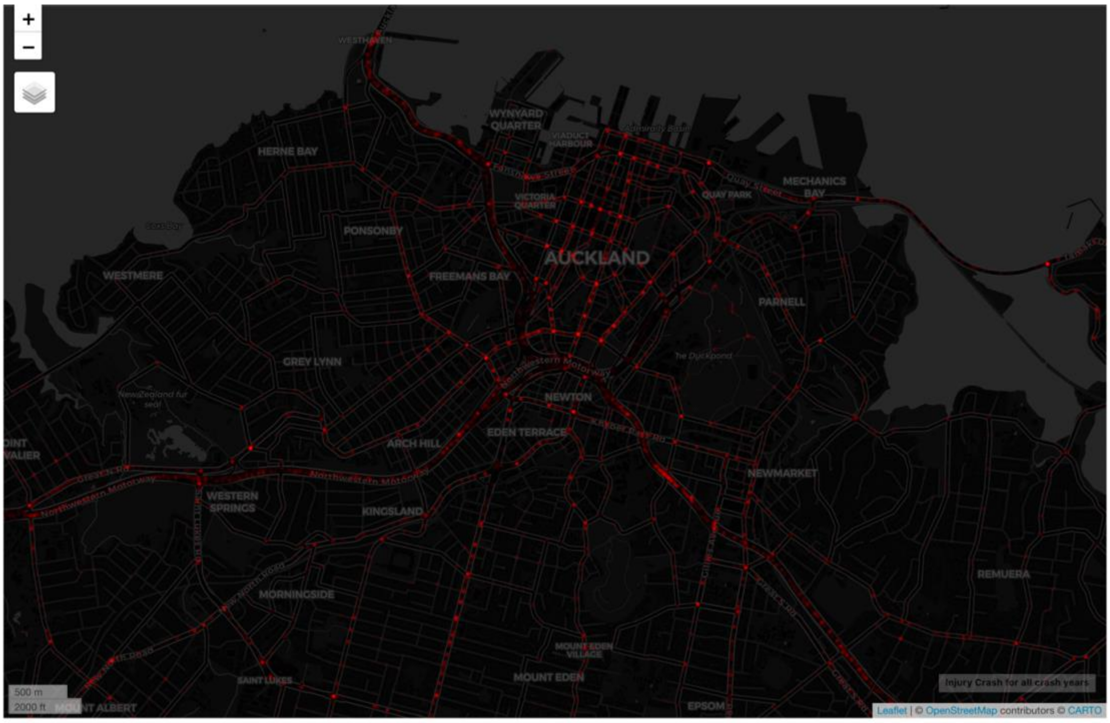


Figure 3.2: Red points represents a single injury crash in Auckland

3.2 Journey to Work Data

The 2018 census's main means of travel to work data is used to derive the journey made by drivers who are 15 years old and over from their usual residence to their place of work through the 2018 census. Data for this study included every unique journey made by a driver of a car, truck or van, either privately owned or company-owned, to estimate the commuter's road safety exposure through the distance it has driven. The usual residence and workplace locations are the centroids of 2018 Statistical Area 2 New Zealand geographic boundaries (SA2) [27]. Auckland Transport, the organisation responsible for all transport services in the Auckland region, has also used this dataset to understand travel patterns. The key findings from their analysis are increasing public transport uptake, increasing walking and cycling in the central suburbs of Auckland and driving being the primary mode of transport to work, especially in the outer urban area [28].

This study focuses on drivers whose usual residence is in the Auckland region (41% of trips, n = 16,035). In figure 3, notice that the usual residence spatial polygons are dispersed in the Auckland region. At the same time, the workplace is concentrated in specific areas such as Penrose, Auckland Airport, Middlemore, Takapuna Central and Parnell West. This means that drivers typically drive from various areas in Auckland to isolated areas for their workplace. The centroids of the SA2 units represent the coordinates of both the usual residence and workplace used to estimate the total distance travelled for each unique trip using the *ghroute* package.

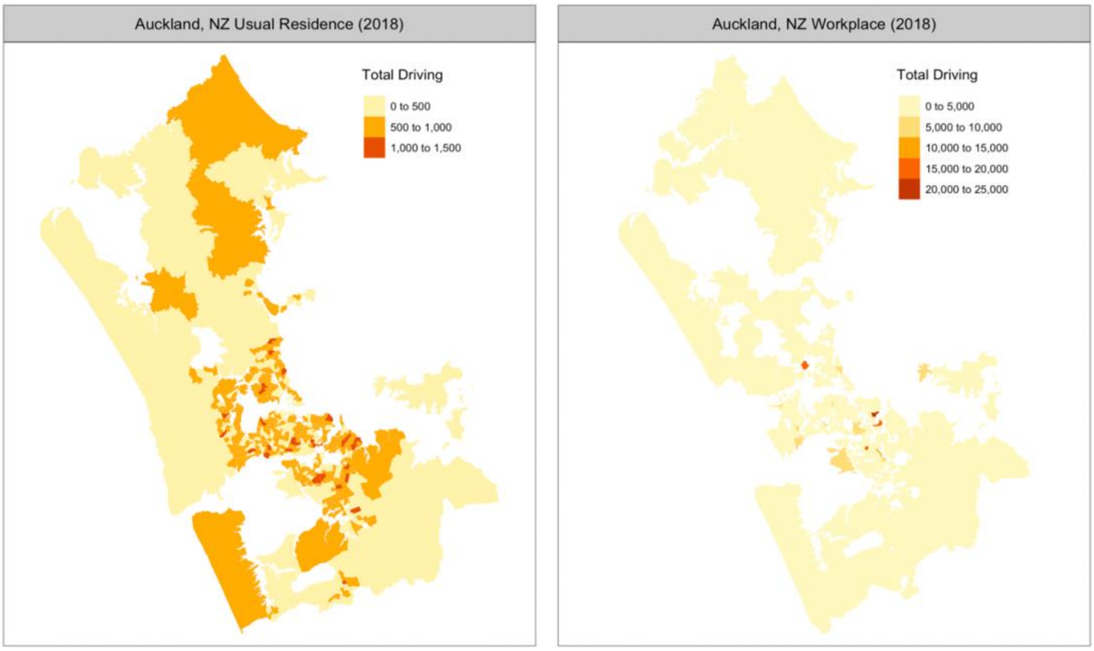


Figure 3.3: Plotting the total number of drivers in their usual residence versus total number of driving per workplace.

The outliers (journey from usual residence to the workplace is greater than or equal to 600 km; 0.01% of all the drivers or 0.02% of all unique car trips) in the dataset are excluded from the analysis. A total of 16,031 trips were made by 293,382 drivers whose usual residence is in Auckland Region.

3.3 Statistical Area 2 Data

The statistical Area 2 clipped polygons are extracted from Statistics New Zealand. It is defined at mesh block and statistical area 1 level geographic units. The statistical area 2 geospatial data are updated over time due to changes in geographic boundaries [29]. There is a total of 2,253 statistical area 2 polygon in the entire New Zealand. 563 of the 2,253 (approximately 25%) are in the Auckland region. In this research, we are using the 2018 statistical area 2 clipped to coastline version to suit the 2018 census main means of work data where the polygons are located in the Auckland region. This data is used to understand the crash volumes for statistical area 2 polygons.

3.4 Auckland Traffic Counts

The Auckland traffic counts dataset is sourced from Auckland Transport. The data is an approximate indication of traffic volumes across different sites in Auckland. The traffic volumes are counted at varying dates and times; for example, AM peak, Mid peak and PM peak, and weekday and weekend volumes. The traffic counts predominantly consider both directions of traffic flow (99.68% of traffic counts in both directions of traffic). The data set also contains information on the coordinates of the exact location of the traffic counters, typically on the side of the road. This data has helped understand traffic volume to support road design, prioritising network improvements, assess road safety risk exposure and determine the effectiveness of past road improvements [30].

In this study, we will use Auckland traffic counts that occurred in 2018 to validate the traffic volume obtained from the simulated routes against actual traffic volumes. As mentioned earlier, the traffic counter is placed on the side of the road. This means it will not intersect with any of the simulated routes. The width of the road varies depending on the type of road and the area. Thus, these need to be accounted for when associating the simulated routes with the traffic volumes from the traffic counter. We will need to determine the optimal distance between the route in either direction of the traffic flow and the location of the traffic counter. Around 30% of the traffic counters are not within a 20 meters radius from the commuter's routes. The main reason for this is that we are using centroids of the statistical area 2 geographic unit rather than the actual residence of the drivers; thus, not all possible routes are simulated.

Chapter 4
Methodology

4.1 Assumptions

The assumptions in this research are that the simulated routes are the best routes from the commuter's usual residence to their respective workplace, the crashes and one-way commute occurred during peak hour (between 7:30 am and 9:30 am) on a weekday and the centroids of the statistical area 2 are the residence of the commuters and workplace. These assumptions are in placed because the datasets available does not have the information we required.

4.2 Overview

The approach in this research involves establishing the necessary road safety measures and methodologies on geospatial analysis. Primarily, it is important that the datasets are prepared and cleaned for analysis. The routes are simulated using the algorithm from *GraphHopper*. This is followed by using the routes to determine the distance travelled exposure. The distance travelled exposure along with the crash data is used to calculate road safety risk. Route Riskiness is also calculated using both crash data and journey to work data. The simulated routes are validated using the data from the traffic counter. In addition, spatial autocorrelation is used to conduct geospatial data analysis.

4.3 Simulating Routes

The *get_routes* function was developed for a *motroadsafety* R package to simulate each unique route utilising the route function from the *ghroute* package. The *ghroute* package simulates a route given a pair of latitude and longitude coordinates of a source and destination locations using routing algorithms from *GraphHopper* via a Java API. An OpenStreetMap (OSM) protocolbuffer binary format (PBF) file of New Zealand is then used to obtain all the possible routes in New Zealand for different modes of travel, i.e., car, public transportation, walking or cycling. This approach contrasts with the *SaferActive* project, which uses the routing algorithm from cyclestreet.net [5].

The methodology used in this research utilises the Simple Features (sf) package for geocomputation. It was developed to adopt the Simple Features hierarchical data model endorsed by the Open Geospatial Consortium (OGC) and a widely supported model that ensures cross-transferrable setups. For example, importing from and exporting to spatial databases. The *sf* package largely supersedes the *sp* package for spatial operations. It has 17 geometry types, for example, points, line and polygons and their respective 'multi' versions. The advantages of using the package are fast reading and writing of data and enhanced plotting performance, moreover, *sf* objects can be treated as data frames [18].

The router must be initialised first before running the *get_route* function by providing the directory of where the OSM PBF file is located and set the profile. In this case, we are using 'car' as the profile as we are interested of commuters driving to work. A four-column numeric matrix is a required input for the *get_routes* function, where values of the matrix are the longitude and latitude coordinates for each source and destination coordinates. The function will return routes of an *sf* data frame where the geometry type is a *linestring*, including the distance travelled of the routes in meters and the time travelled in seconds. For source and destination pairs where no route is possible, the function will return the sf*_empty* geometry type. The *get_routes* function also has the option to transform the resulting geographic coordinates to any coordinate reference system as long as it provides its target coordinate reference system either by giving an EPSG code or a proj4string definition.

The 2018 main means of travel to work data only contains projected coordinates of the usual residence and workplace locations. Thus, the projected coordinates are transformed to geographic coordinates, i.e., latitude and longitude, using the *sf_transform* function in the sf package to comply with the constraints of the *get_route* function.

4.4 Exposure

The *get_dist_travelled* function from the *motroadsafety* R package is used to calculate the exposure aggregated by polygon geometry. Exposure is defined as a road safety measure of the amount travelled by a commuter [12], in this research we are proposing to use distance travelled by the commuter. The required input results from the *get_routes* functions or any data frame containing an *sf linestring* geometry and a data frame with *sf polygon* geometry. Weight is an optional argument in the *get_dist_travelled* function. It represents the total number of commuters travelling on a given route.

The function utilises the *st_intersection* function from the sf package to spatially clipped the routes into line segments that is a spatial intersection of provided polygons. The function's output returns an sf data frame containing the input polygons and the aggregated total distance measured in meters derived from the length of the clipped route *sf linestrings*. We have used the results from the *get_routes* function and statistical area 2 polygons as inputs to the *get_dist_travelled* function to return the exposure of each statistical area 2 polygon. The results were visualised using the *mapview* R package.

4.5 Road Safety Risk

The *get_risk* function was the method developed to calculate the road safety risk for a given geographical unit. Merlin, Guerra and Dumbaugh [16] defined risk as to the probability of a crash per unit exposure. In this study, we calculate risk as the total number of crashes within the graphical unit per total distance travelled within the same geographic unit. The *get_risk* function requires crash coordinates data frame with *sf point* geometry type and exposure data, or the distance travelled by the commuter aggregated by a geographical unit. Geospatial data, crash data frame and exposure data frame, must have a projected CRS. Crash weight, buffer and CRS are optional arguments. Crash weight refers to the total number of injuries during a crash. Buffer refers to extending the polygon geographic unit by a particular unit of measure. This is because the statistical area 2 polygons lie between roads, so a specific crash site can fall between more than one geographic unit. The default buffer is 10 meters, but this is arbitrary. CRS is an optional argument that will transform the function's output CRS to the CRS provided by the user.

The *get_risk* function uses the extended polygon using the buffer argument and joins it with the crash data. The total crash weight is calculated by aggregating the crash weight by exposure data. The risk is then calculated by dividing the total crash weight against the exposure, i.e., the total distance travelled in meters derived from the *get_dist_travelled* function. The *get_risk* function returns a data frame with an sf column containing a polygon geometry type. The output data frame also includes a new column called 'risk', which is the road safety measure where the unit is dependent on the input data CRS. In this study, the output from the *get_dist_travelled* and the CAS data coordinates, along with the sum of the serious injury and fatal count, is used as the crash weight. The geometries of the input data frames are projected by coordinated reference systems of EPSG code 2193. The resulting output is then used to analyse the road safety risk for each statistical area 2 geographic unit.

4.6 Route Riskiness

The *route_risk* function is the method developed to calculate the risk for each unique source and destination pair. The function required the user to provide two sf data frames, the routes of the data frame containing the route of line string geometry type and the crash of data frame containing crash sites as sf points geometry type. Both data frame arguments with sf columns require a projected coordinate reference system. Radius is an optional argument determining the total number of crashes from a particular route. The default radius is 5 meters, an arbitrary distance. The *route_risk* function takes individual routes and aggregates the total number of crashes within a 5 meters radius from the said route. It returns the routes of the data frame with line string geometry and includes a new column *crash_within_5m*, the aggregated crash sites encountered by a given route.

The output from *get_routes* functions is used as the routes argument, and crash data is the CAS data transformed into an sf data frame with an sf column of point geometry type. The output of the *risk_routes* function is used to analyse the relationship between the total crash encountered by a given commuter and the distance travelled by the route in meters or route riskiness. A linear model was used to describe the relationship between the total crash encounter and the distance travelled.

4.7 Validating Routes

The routes simulated are from census data extracted by Statistical New Zealand a sample of the actual population. We wanted to quantify the difference between the total traffic volumes from the route simulated against the approximate traffic volumes from the traffic counter collected by Auckland Transport to measure the performance of the route estimated. This is a similar approach conducted by the SaferActive project, where researchers validate the cycling and traffic counts [5]. It is essential to highlight the caveat with the source and destination pairs of every unique trip from the 2018 main means of travel to work data have been aggregated to statistical area 2 geographic units to preserve sensitive personal data of the survey respondents. The routes from the *get_routes* function represent the total volume by multiplying the total drivers on a given route that is close proximity to the traffic counter. Since the traffic counter is typically not located on the road and road widths vary depending on the road and area thus, we had to consider that the traffic counter intersects with the simulated route. This is achieved using the *st_within_distance* from the sf package to evaluate the optimal distance between a given route and the Auckland traffic counter. We tested for various lengths, 10 meters, 15 meters and 20 meters. The traffic volumes from the simulated routes and Auckland traffic counter are then compared with Auckland traffic counter volumes (for varying time) after the optimal distance between the traffic counter and routes have been established. This analysis will enable us to validate the traffic volumes from the simulated routes.

4.8 Spatial Autocorrelation

The *spdep* R package is used to conduct exploratory data analysis [20], specifically spatial autocorrelation for hotspot detection [22]. The variable of interest in this research is the total number of crashes encountered by a given driver aggregated to its usual residence spatial unit. Spatial autocorrelation will enable us to determine if there is spatial dispersion globally and locally on the route riskiness.

The first step is to obtain the spatial weights of each statistical area 2 polygon and its distance from its neighbours. However, there are polygons in Auckland that have no neighbours because they are on an island. Thus, a cleaning step must be performed to remove the islands with no neighbours. The *ms_filter_islands* function from the *rmapshaper* R package was used to remove the islands. The spatial data is sufficiently clean enough to calculate the spatial weights using the *spdep* package *poly2nb* function with the queen argument set to False because more than one point is shared. After obtaining the spatial weight, *Moran's I* test was conducted to determine that global spatial autocorrelation for risky routes is statistically significant. Once global spatial autocorrelation was confirmed, we proceeded to Local Spatial Autocorrelation (LISA) to detect local clusters where the polygons are categorised into four categories – High-high, High-low, Low-High and Low-low. The *localmoran* function from the *spdep* package was used to calculate LISA.

Chapter 5
Results

5.1 Simulating Routes

The 2018 census main means to work data where the usual residence of the driver is in the Auckland region had 16,031 unique routes. The *get_route* function was only able to simulate 15,929 unique routes. There were 102 unique usual residence and workplace pairs that had no routes. South Head, Kawakawa Bay-Oreore, Waipatukahu, Waitakere Ranges South, Waitakere Ranges North and Okura Bush are the predominant usual residence statistical area 2 geographic units that had no possible route between drivers' usual residence to their workplace. Another issue we encountered is that not all busy roads in Auckland have been exhausted. For example, drivers usually residing in Mission Bay and Orakei East tend to take Kepa Road and Kohimarama road rather than Tamaki drive which is be particularly closer for some of the houses in these areas. These issues can be attributed to the fact that the source and destination coordinates were derived from the centroids of the statistical area 2 geographic units.

In figure 5.1, simulated routes were visualise using the *mapview* R package. It is noticeable that the routes simulated from the *GraphHopper* algorithm is predominantly taking the motorway, hence high density on the southern, northern, northwestern, and southwestern motorways of Auckland. Meanwhile, local streets like Tamaki Drive are less dense. The results from this step will be used as an input to calculating exposure per SA2 polygon.

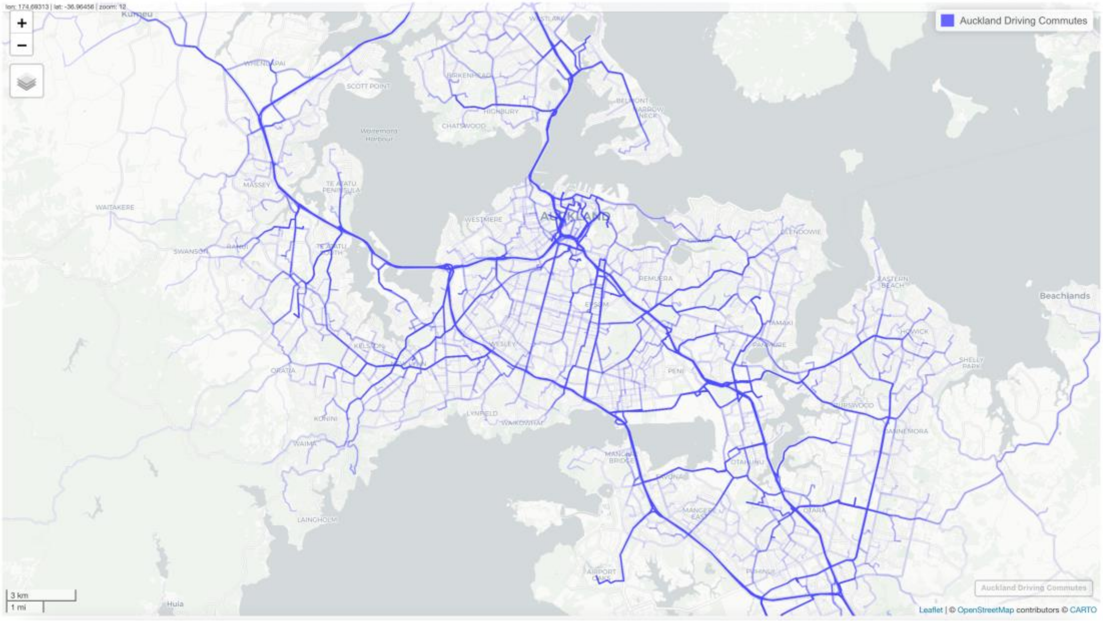

Auckland Driving Commutes
AUCKLAND
WAITAKERE
Beachlands
3 km
1 mi
Leaflet | © OpenStreetMap contributors © CARTO

Figure 5.1: A plot of all the routes simulated from *get_route* function

5.2 Exposure

The exposure in this research is expressed as distance travelled, similar to the approach taken by the *SaferActive* project [5]. However, unlike the SaferActive project the distance travelled exposure is represented daily through taking into consideration the total number of working days. In this study, we have annualized distance travelled exposure.

Here we have obtained the total distance travelled by each driver that traversed for each SA2 polygon. In figure 5.2, visualizes the aggregated distance travelled for each travelled by SA2 polygon using the *tmap* package. We observed that the SA2 areas workplace is seemingly correlated with the total drivers aggregated to SA2. Indicating that commuters tend to travel similar areas for work. For example, Penrose and Auckland Airport are highly dense for both the total count of drivers in SA2 workplace and the total distance travelled per SA2. The result in this step is subsequently used to calculate road safety risk.

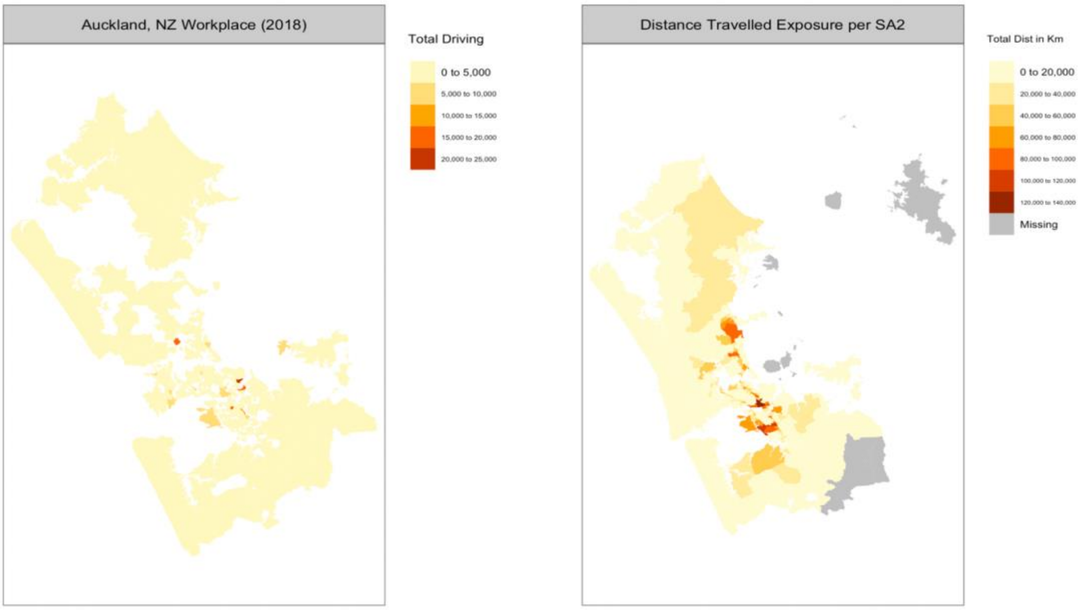


Figure 5.2: Total count of drivers aggregated by workplace Statistical Area 2 polygon (left). Total distance travelled aggregated by Statistical Area 2 (right)

5.3 Road Safety Risk

The road safety risk was calculated using the *get_risk* function from the *motroadsafety* package. The R package *mapview* was used to visualise the risk for various SA2 polygons, refer to figure 5.3. It is noticeable in figure 5.3 that SA2 polygons where the commuters travel less such as Kawakawa Bay-Orere, South Head, Rowandale West, Burbank and Clendon Park West but were exposed to either a fatal or serious crash occurred are the riskier in comparison to areas where commuters are more likely to travel such as the New Market, Grey Lynn, Central Mount Wellington and Manukau Central and had considerable crashes. Accordingly, areas where there is less traffic volume, but high count of traffic accidents tend to be riskier compared to areas where there is high traffic volume and comparable traffic accidents.

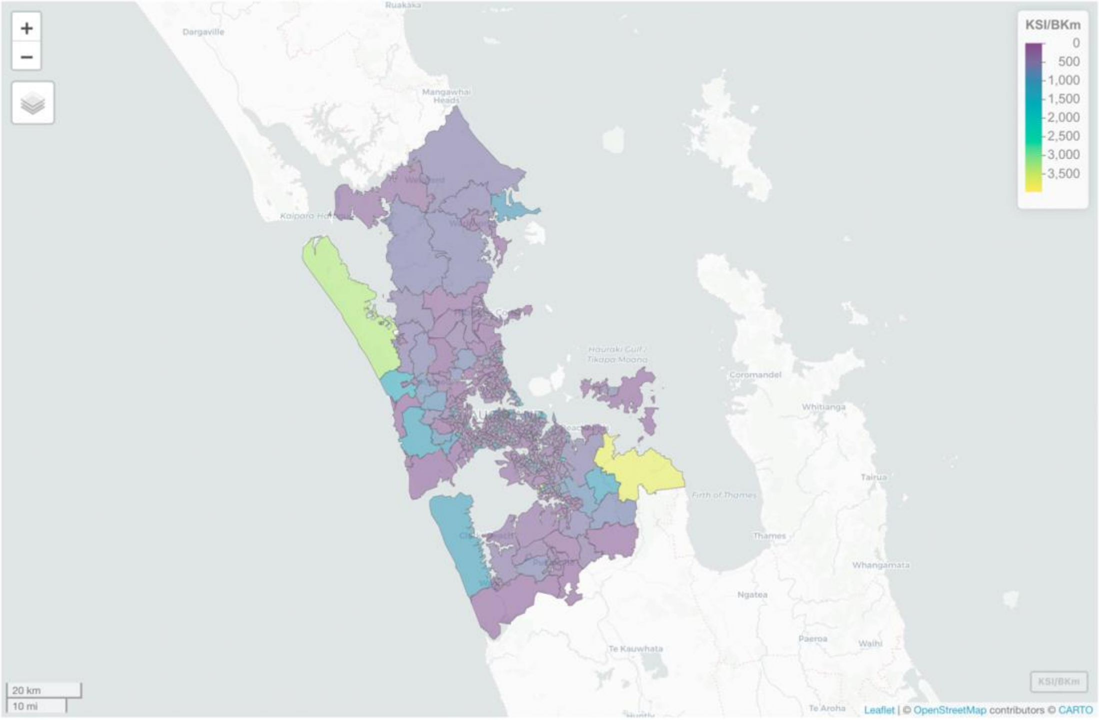


Figure 5.3: 2018 Risk (KSI per BKm) per Statistical Area 2 (SA2)

5.4 Route Riskiness

Route riskiness enables us to quantify the riskiness of a commuter's journey from their usual residence to the driver's workplace through the total number of crashes it has encountered. In figure 5.4 shows a plot of one route travelling from its usual residence in Pukekohe Central to its workplace in Shortland Street which is in Auckland Central. The 9 drivers taking this approximately 54 Km route mostly drove in the southern motorway has encountered 431 crashes which is located within 5 meters radius from the route in 2018. These crashes includes both non-injury and injury crashes.

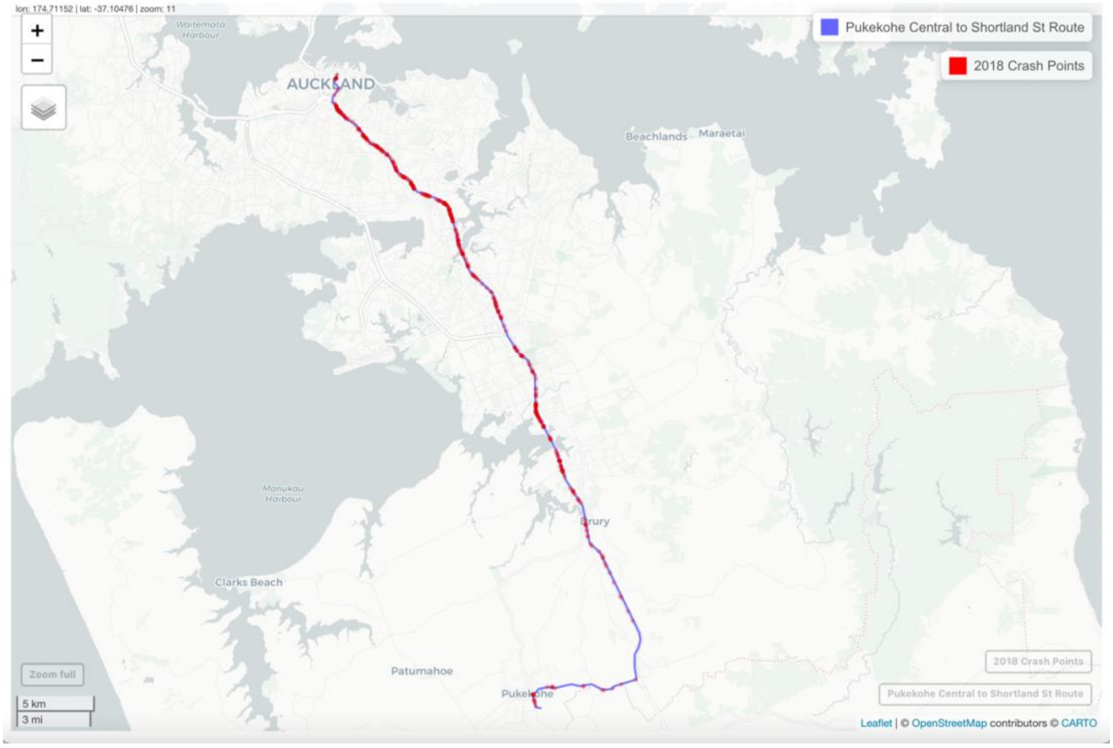

lon: 174.71152 | lat: -37.10476 | zoom: 11
+
−
Waitematā Harbour
AUCKLAND
Pukekohe Central to Shortland St Route
2018 Crash Points
Beachlands
Maraetai
Manukau Harbour
Drury
Clarks Beach
Patumahoe
Pukekohe
Zoom full
5 km
3 mi
2018 Crash Points
Pukekohe Central to Shortland St Route
Leaflet | © OpenStreetMap contributors © CARTO

Figure 5.4: Plot of a risky route from usual residence to workplace

We fitted a simple linear regression to model relationship between the total count of crashes encountered by the driver within 5 meters from the route and the total distance travelled by the driver (refer to figure 5.5). In figure 5.5, we can see that there is an increasing relationship between total crashes and total distance travelled. The residual plot of a simple linear (refer to figure 5.6) indicated a strong curvature trending downwards and the scatter plot does not have a constant scatter. The model explained 54% of the variability in the total count of crashes encountered. In reference to the research conducted by Pei et al. (2012), the occurrence of collision is positively correlated to distance travelled exposure with consideration of the speed. In this study, we have not taken into account the effects of speed on the probability of a crash due to complexity. This needs to be taken into consideration in future work. Moreover, there are also crashes that occurred at short distances, and this could be due to the 'close to home' effect as established by Burdett et al. where drivers tend to be complacent because it has become a routine route [17].

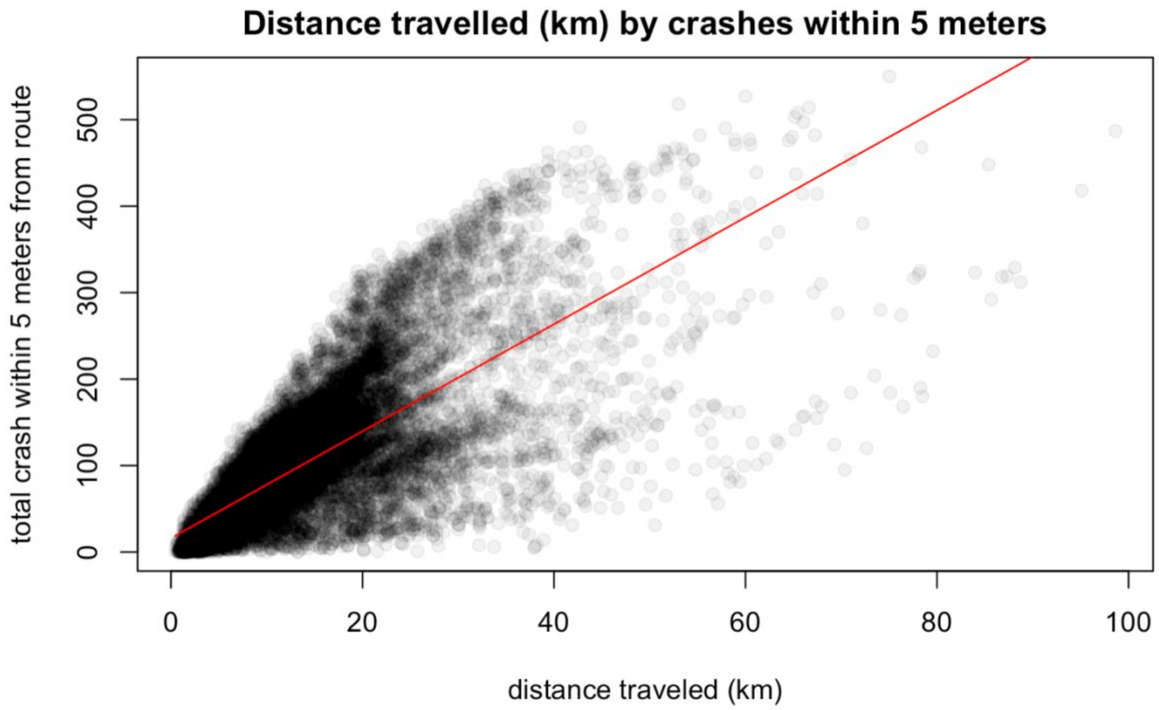


Figure 5.5: Distance traveled (km) by total crashes within 5 meters from route

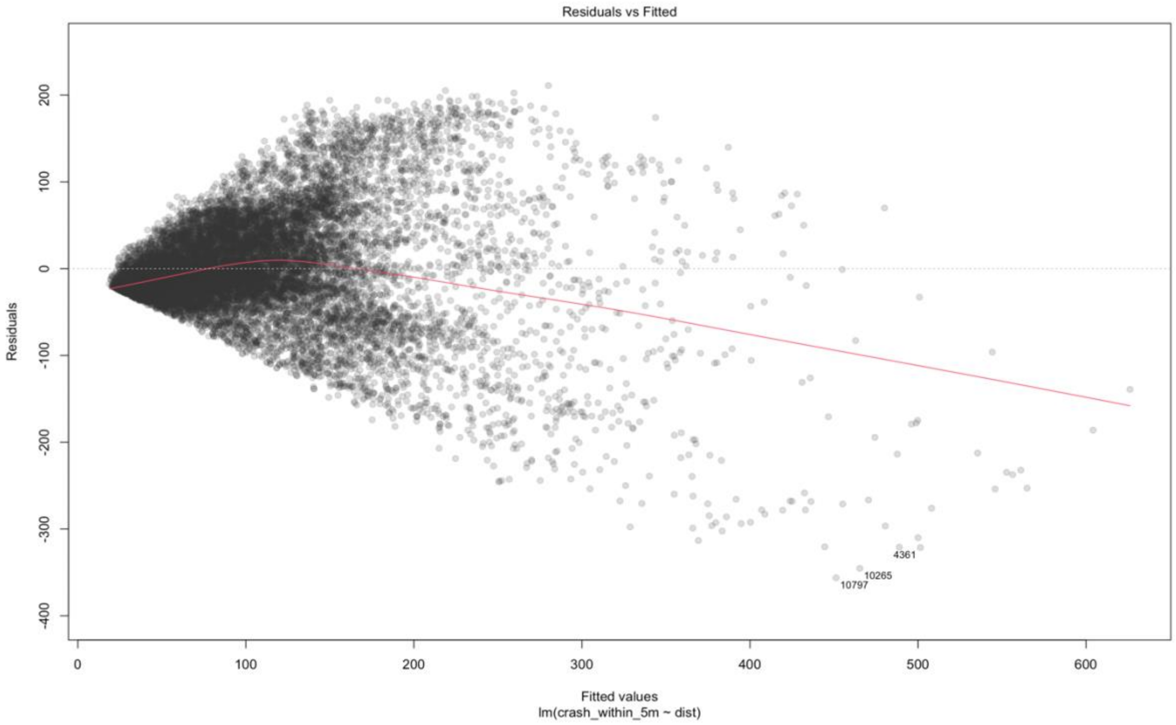


Figure 5.6: Residual plot

5.5 Validating Routes

Validating traffic as mentioned involved determining the optimal distance between the simulated routes and traffic counter assuming that the traffic counter can be either side of the road and that at some radius the traffic counter is going to intersect with the route that has traffic flows on both directions.

In figure 5.7, we have determined that 20-meters is the optimal distance between the commuter route and traffic counter, after comparing it with 10- and 15-meters distance. Approximately 30% (651 of 2331) traffic counters are not within 20 meters radius of a given route. This is mainly attributed to the fact that our source and destination are centroids of SA2 polygons, which means that we might have not all possible routes in Auckland (refer to figure 5.8).

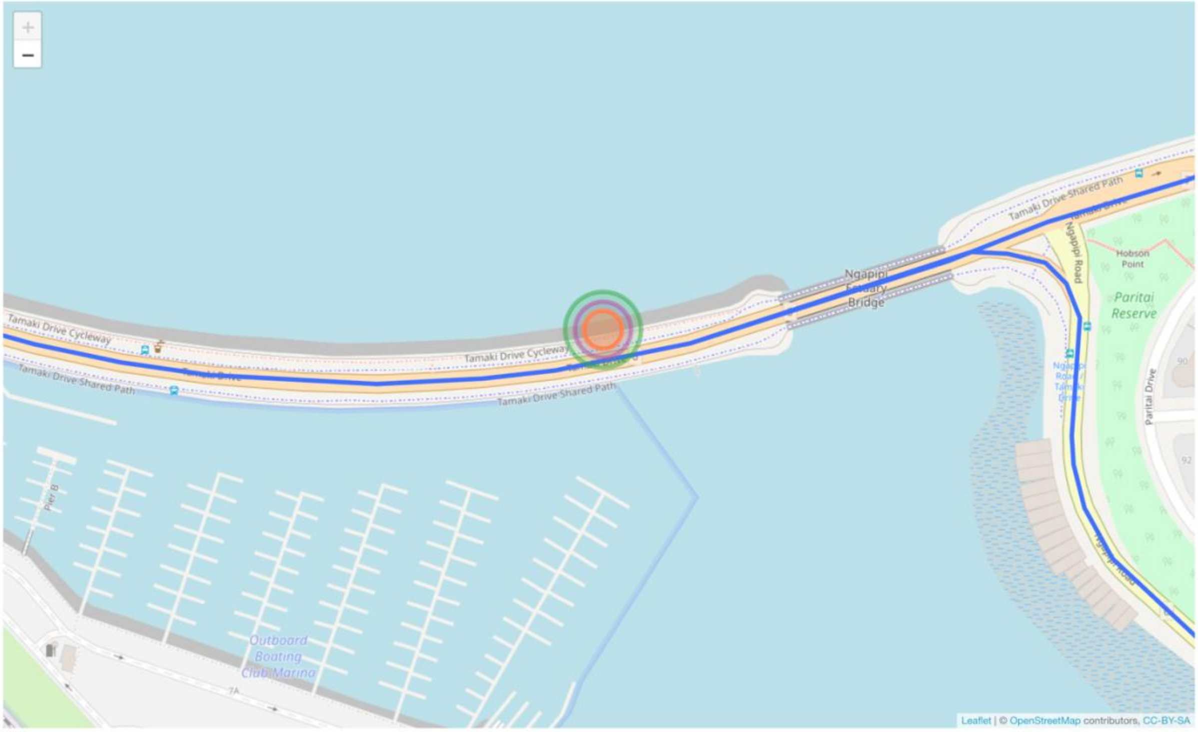


Figure 5.7: Traffic Counter distance from the route. *Red* refers to the traffic counter. *Yellow* refers to traffic counter with 10 meters radius, *purple* for 15 meters radius and *green* for 20 meters radius

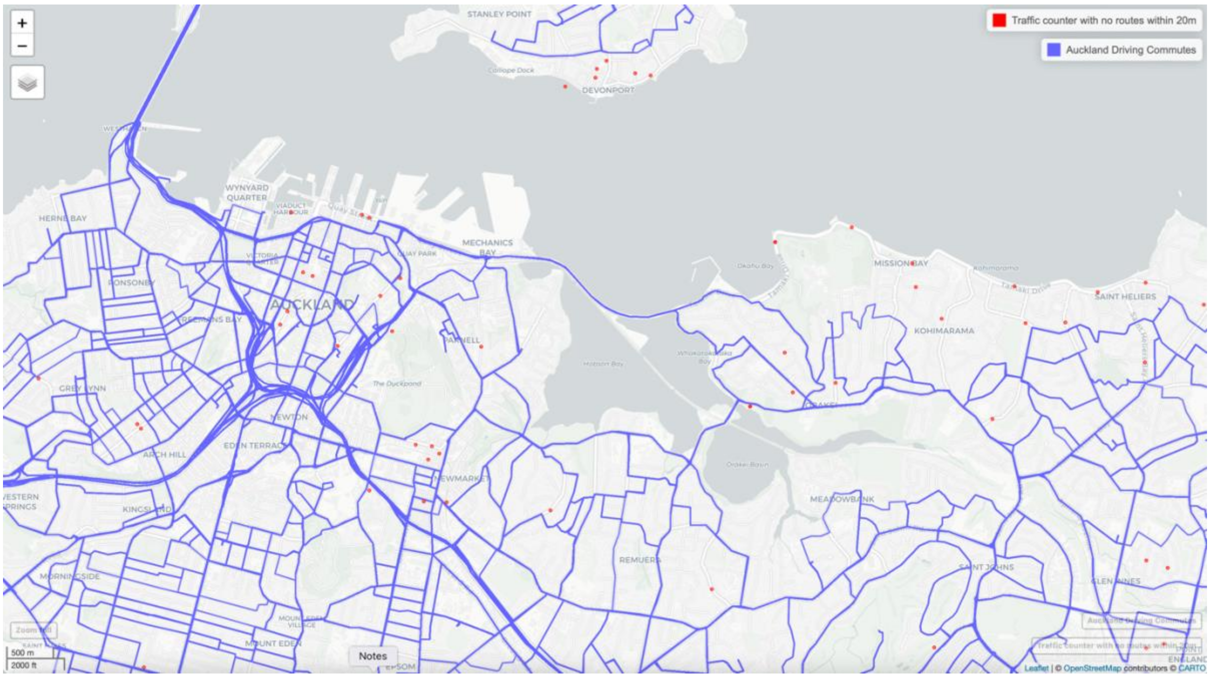


Figure 5.8: Traffic counter (red dots) versus commuter routes (blue lines) in Auckland

After establishing the optimal distance, we plotted the relationship between the simulated total traffic volumes and the total sum of the traffic AM Peak Volume and Mid Peak Volumes during weekday using a scatter plot (refer to figure 5.9). The mean absolute percentage error between the total simulated traffic volume and the sum of the AM Peak Volume and Mid Peak Volumes is approximate 80%. Figure 5.9 indicates that our estimate of traffic volumes is imprecise. This can be attributed to the fact that we are using SA2 centroids to simulate the routes rather than actual residence and other confounding factors such as the journey to work data is a sample data of the Auckland population. For example, the traffic counter will have included all journeys, including public transportation and journeys other than work that is not within the scope of this study.

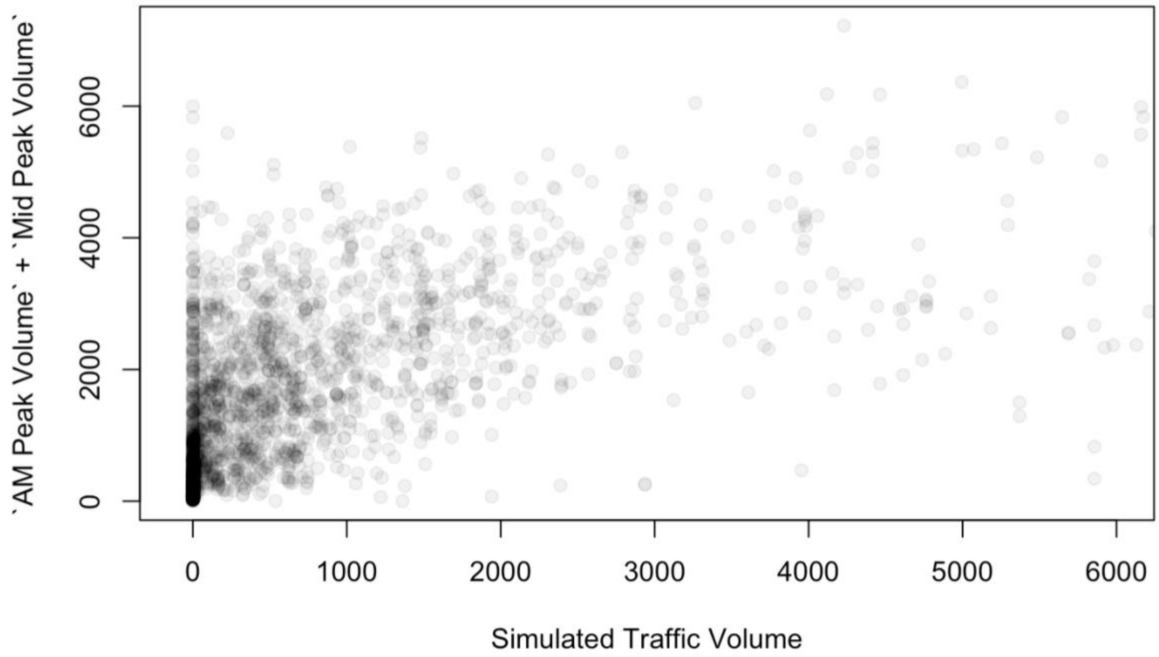


Figure 5.9: Simulated Traffic Volume versus the sum of AM and PM Peak Traffic Volume

5.6 Spatial Autocorrelation

The Moran I statistics is 0.62, and the p-value has a significantly low value. Thus, we can establish that the total crash sites encountered by the driver from its SA2 usual residence are positively autocorrelated and statistically significant. Figure 5.10 coincides with Moran's I statistics indicating positively correlated observations by the riskiness of commuters' journey from their usual residence. This refers to the presence of spatial clusters globally.

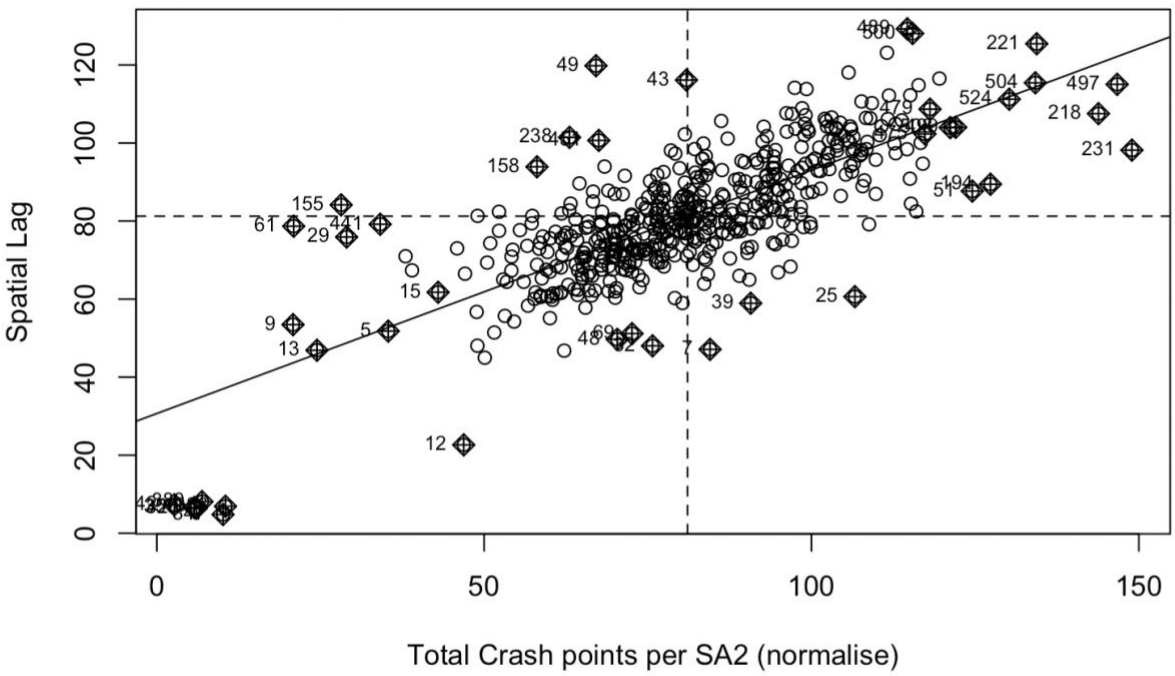


Figure 5.10: Moran's Plot for Global Spatial Autocorrelation

The Local Indicators of Spatial Associate (LISA) map in figure 5.11 shows the variations of risky routes from a SA2 usual residence. The bold red colors indicate neighbors clustered together which have encountered significant riskier routes surrounded by significant riskier routes as well. These clusters are mostly located in South of Auckland where driving on motorways is more likely, and the distance travelled is longer. There are also solid red clusters in West of Auckland, Titirangi areas. The dark blue areas indicate route risk are low from the driver's SA2 usual residence, also surrounded by areas with low route risk are spread across Auckland, New Zealand. In addition, Low-High and High-Low areas where high and low risk areas are next to each other are also spread across Auckland, New Zealand.

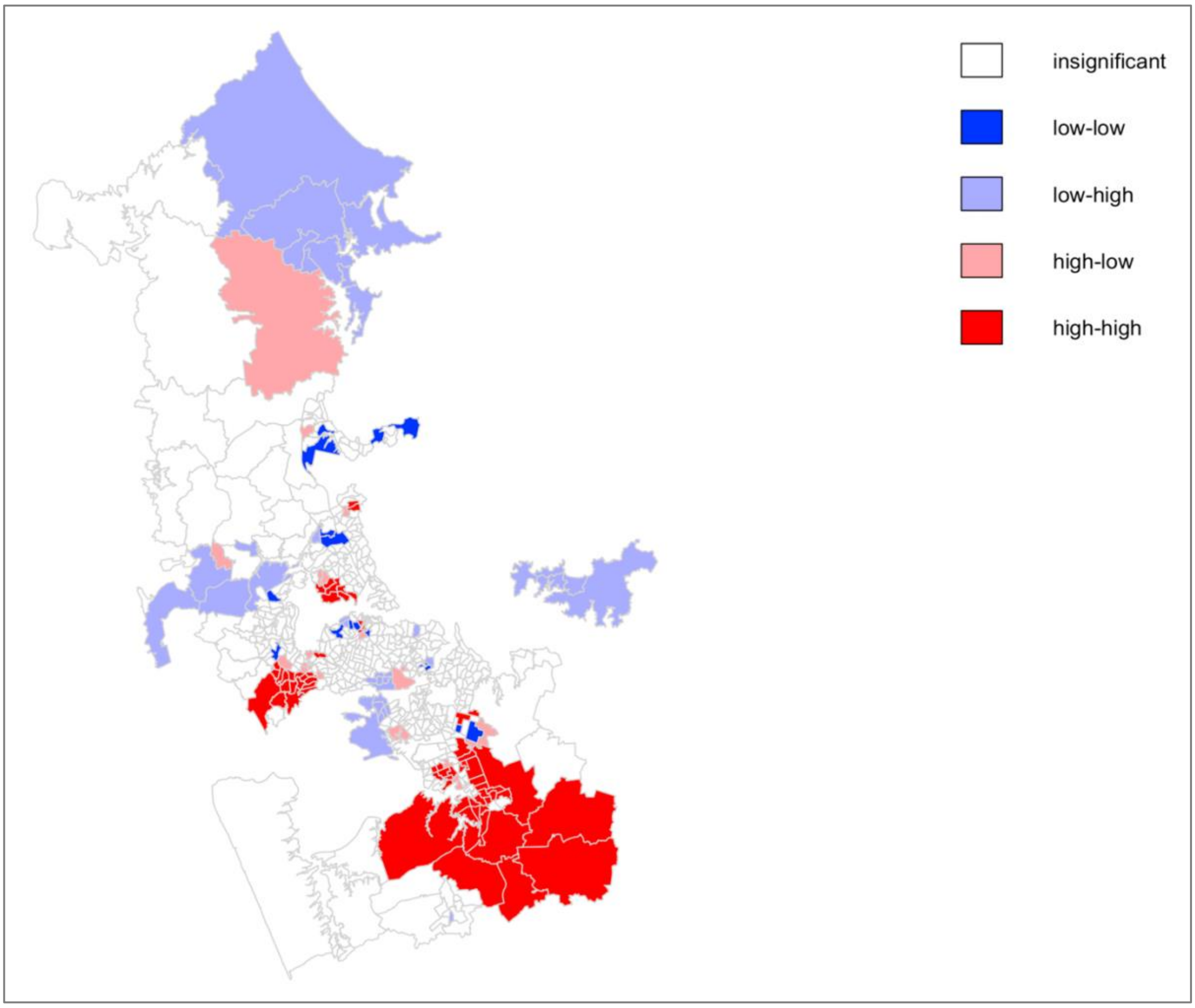


Figure 5.11: Local Indicators of Spatial Associate (LISA) Plot

5.7 Summary

In this chapter, we have presented the results of the geospatial data analysis using the datasets and functions bundled in the developed motroadsafety R package. We have shown that the developed R package can be used to analyse road safety in Auckland, New Zealand citing the relationship between the distance travelled exposure and the total crashes such that the longer the distance travelled indicates the higher the likelihood that a commuter may encounter a crash. Moreover, there are cases where crashes during short distances occurred because of the close to home effect. In addition, we were able to determine the hot spots and cold spots based on the simulated routes. However, discrepancy between traffic volume from the simulated routes and traffic countered requires additional work due to underlying factors.

Chapter 6
R Package Overview

The *motroadsafety* R Package developed in this research bundles together the data and functions used in spatial data analysis on road safety. It is accessible on Github. The benefits of bundling this work in an R package are that it is easy to share the work carried out, reproducible data analysis, and iterate and evolve the work [4].

The *motroadsafety* R package can be installed from Github through below command.

```
Devtools::install_github(“kathycong/motroadsafety”)
```

Other than *motroadsafety* package, *dplyr* and *ghroute* are essential R packages in spatial data analysis in this research. These packages can be loaded in R using below commands.

```
Library(motroadsafety)
library(dplyr)
library(ghroute)
```

The R package has three datasets, *cas_data_akl_2018_proj*, *jtw_driving_akl,* and *sa2_clipped_proj_akl.* The *cas_data_akl_2018_proj* contains data on crashes in Auckland, New Zealand that occurred in 2018. The *jtw_driving_akl* contains data on the commuters' coordinates. The *sa2_clipped_proj_akl* contains data on the Statistical Area 2 polygons. These datasets are accessible using the below commands.

```
Data(“cas_data_akl_2018_proj”)
data(“jtw_driving_akl”)
data(“sa2_clipped_proj_akl”)
```

The *get_routes* function was used in this study to simulate the routes for each commuter. It is essential that the router must be initialised by using the below command before using *get_routes*. The router has the profile argument to set the mode of transport. The default value for the profile command is ‘car’. The router function also requires the file directory of *osm pbf* file that will be used to simulate the routes in New Zealand.

```
## initialising the router for New Zealand
ghroute::router(osm.file = “osm/new-zealand-latest.osm.pbf”)
```

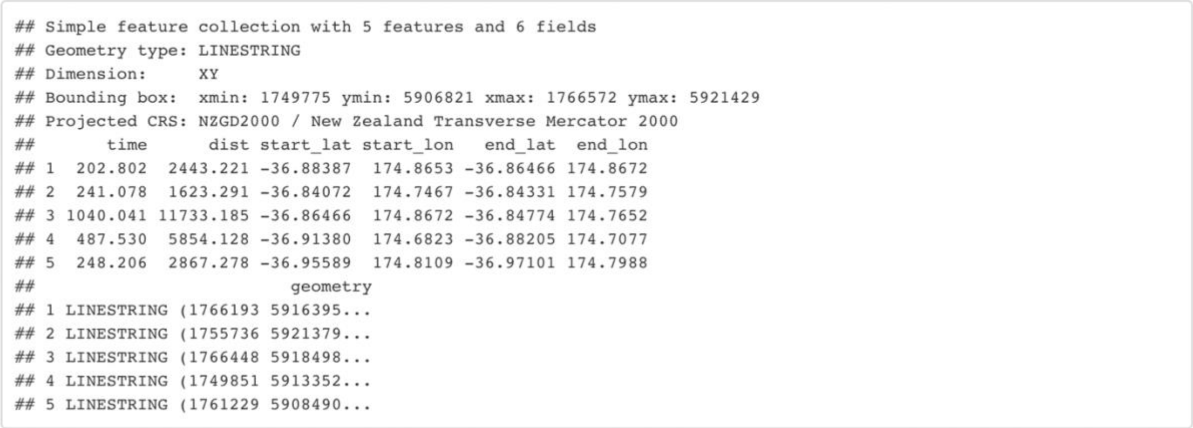
The *get_routes* function requires a matrix or an array with 4 columns. In this dissertation we have transformed the *jtw_driving_akl* data frame into a matrix before passing it through the *get_routes* function, refer to below command.

```
## transform jtw_driving_akl df into a matrix
m <- as.matrix(jtw_driving_akl[c(“start_lat”, “start_lon”, “end_lat”,
“end lon”)])
```

The *m* matrix from the previous command which contains start latitude, start longitude, end latitude and end longitude values of the driver’s journey was used to simulate the routes for each driver and set the options argument CRS to 2193 which is the EPSG code for the typical projection used in New Zealand geospatial data.

```
## getting the routes
jtw_proj <- get_routes(m, crs = 2193)
```

The *get_routes* function will return a data frame with the same length as the input matrix or array with an sf column containing an *sf linestring* geometry type as seen below. The function will also include a time column which represents the time travelled in seconds and *dist* column which represents the distance travelled in meters.



```
## Simple feature collection with 5 features and 6 fields
## Geometry type: LINESTRING
## Dimension:     XY
## Bounding box:  xmin: 1749775 ymin: 5906821 xmax: 1766572 ymax: 5921429
## Projected CRS: NZGD2000 / New Zealand Transverse Mercator 2000
##       time      dist start_lat start_lon   end_lat  end_lon
## 1  202.802  2443.221 -36.88387  174.8653 -36.86466 174.8672
## 2  241.078  1623.291 -36.84072  174.7467 -36.84331 174.7579
## 3 1040.041 11733.185 -36.86466  174.8672 -36.84774 174.7652
## 4  487.530  5854.128 -36.91380  174.6823 -36.88205 174.7077
## 5  248.206  2867.278 -36.95589  174.8109 -36.97101 174.7988
##                        geometry
## 1 LINESTRING (1766193 5916395...
## 2 LINESTRING (1755736 5921379...
## 3 LINESTRING (1766448 5918498...
## 4 LINESTRING (1749851 5913352...
## 5 LINESTRING (1761229 5908490...
```

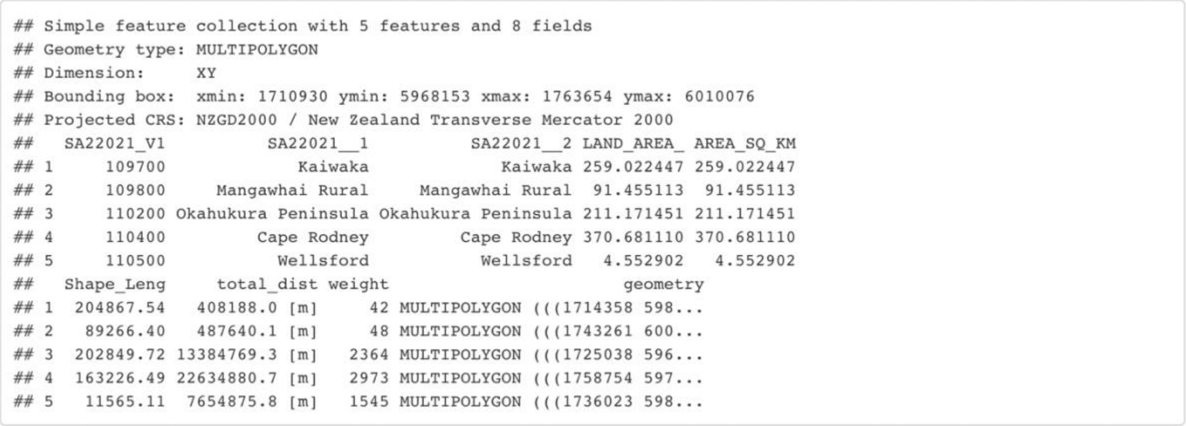
The result of the *get_routes* function is enriched with the data from the *jtw_driving_akl* which contains other information, e.g., total number of drivers in a given route, SA2 polygon names of the usual residence and workplace and etc, using the *cbind* function. The variable *jtw_proj* is column bind into the *jtw_driving_akl* and *jtw_proj* data frames.

```
## getting distance travelled per SA2 polygon
jtw_proj <- cbind(jtw_proj, jtw_driving_akl)
```

The *get_dist_travel* function in the *motroadsafety* package calculates the distance travelled exposure per polygon. It requires two arguments, 'polygon', a data frame containing an *sf polygon* geometry type, and 'routes', a data frame containing an *sf linestring* geometry type. The distance travelled per polygon is calculated using the *sf intersection* function that determines the line segments of the routes that are the intersection of the provided polygon. The weight argument is optional. It is the number of commuters for each route in this research. The code below used sa2_clipped_proj_akl as the polygons, *jtw_proj* as the routes and *jtw_proj$total_driving* vector as the weight.

```
## getting distance travelled per SA2 polygon
akl_dist_travel <- get_dist_travel(sa2_clipped_proj_akl,
                                   jtw_proj,
                                   weight = jtw_proj$total_driving)
```

The *get_dist_travel* function will return a data frame with columns from the polygon data frame and new columns *total_dist* and *weight*, which is the total length of line segments in a polygon. The unit of measure is dependent on the polygon CRS provided but in this research the resulting unit is meters. The weight is an aggregated sum of commuters per polygon. The resulting geometry type is the same as the polygon argument.

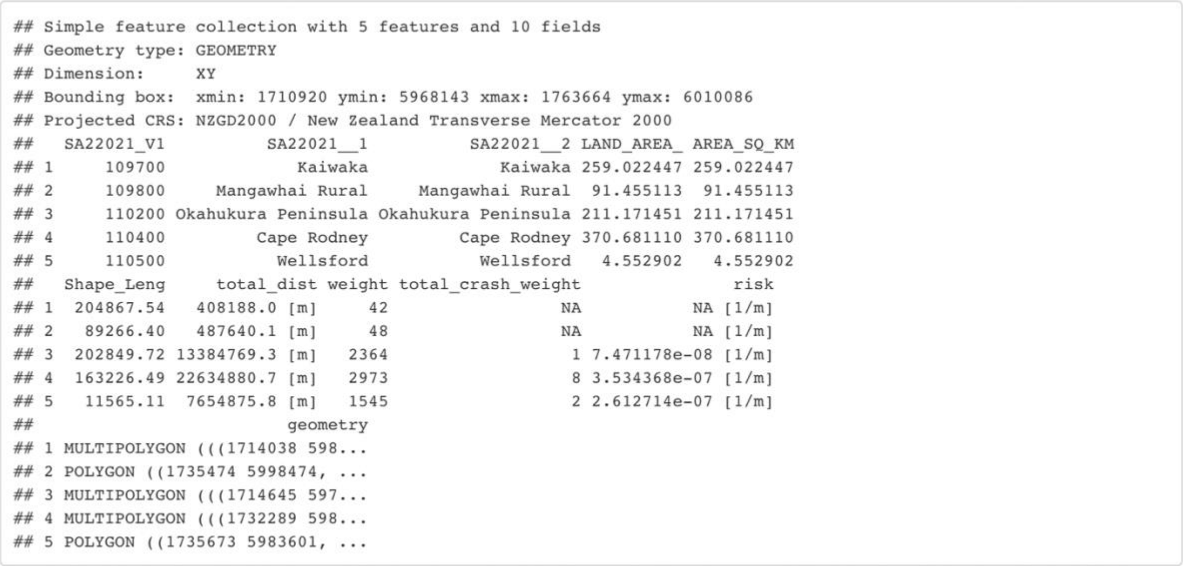


```
## Simple feature collection with 5 features and 8 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: 1710930 ymin: 5968153 xmax: 1763654 ymax: 6010076
## Projected CRS: NZGD2000 / New Zealand Transverse Mercator 2000
##   SA22021_V1          SA22021__1          SA22021__2 LAND_AREA_ AREA_SQ_KM
## 1     109700             Kaiwaka             Kaiwaka 259.022447 259.022447
## 2     109800     Mangawhai Rural     Mangawhai Rural  91.455113  91.455113
## 3     110200 Okahukura Peninsula Okahukura Peninsula 211.171451 211.171451
## 4     110400        Cape Rodney         Cape Rodney 370.681110 370.681110
## 5     110500          Wellsford           Wellsford   4.552902   4.552902
##   Shape_Leng     total_dist weight                       geometry
## 1  204867.54   408188.0 [m]     42 MULTIPOLYGON (((1714358 598...
## 2   89266.40   487640.1 [m]     48 MULTIPOLYGON (((1743261 600...
## 3  202849.72 13384769.3 [m]   2364 MULTIPOLYGON (((1725038 596...
## 4  163226.49 22634880.7 [m]   2973 MULTIPOLYGON (((1758754 597...
## 5   11565.11  7654875.8 [m]   1545 MULTIPOLYGON (((1736023 598...
```

The *get_risk* function is used to calculate the road safety risk by providing three arguments, the crash data, weight and distance travelled exposure. The crash data is a data frame with an *sf point* geometry type. The *cas_data_akl_2018_proj* is used as the crash data. The sum of the serious and fatal injuries from *cas_data_akl_2018_proj* data frame is used as the weight and the distance travelled exposure was derived from the *get_dist_travel* function (refer to below command).

```
## getting the risk
risk_proj <-  get_risk(cas_data_akl_2018_proj,
                       cas_data_akl_2018_proj$seriousInjuryCount + ca
s_data_akl_2018_proj$fatalCount,
                       akl_dist_travel)
```

The result of the *get_risk* function is assigned to the variable *risk_proj*. The function returns a data frame with an *sf polygon* geometry type and a new column called ‘risk’, which is the total crash weight divided by the total distance travelled, where the unit of measure is dependent on the unit of measure provided but in this research the unit of measure is meters. Below is a sample of the output from the *get_risk* function.

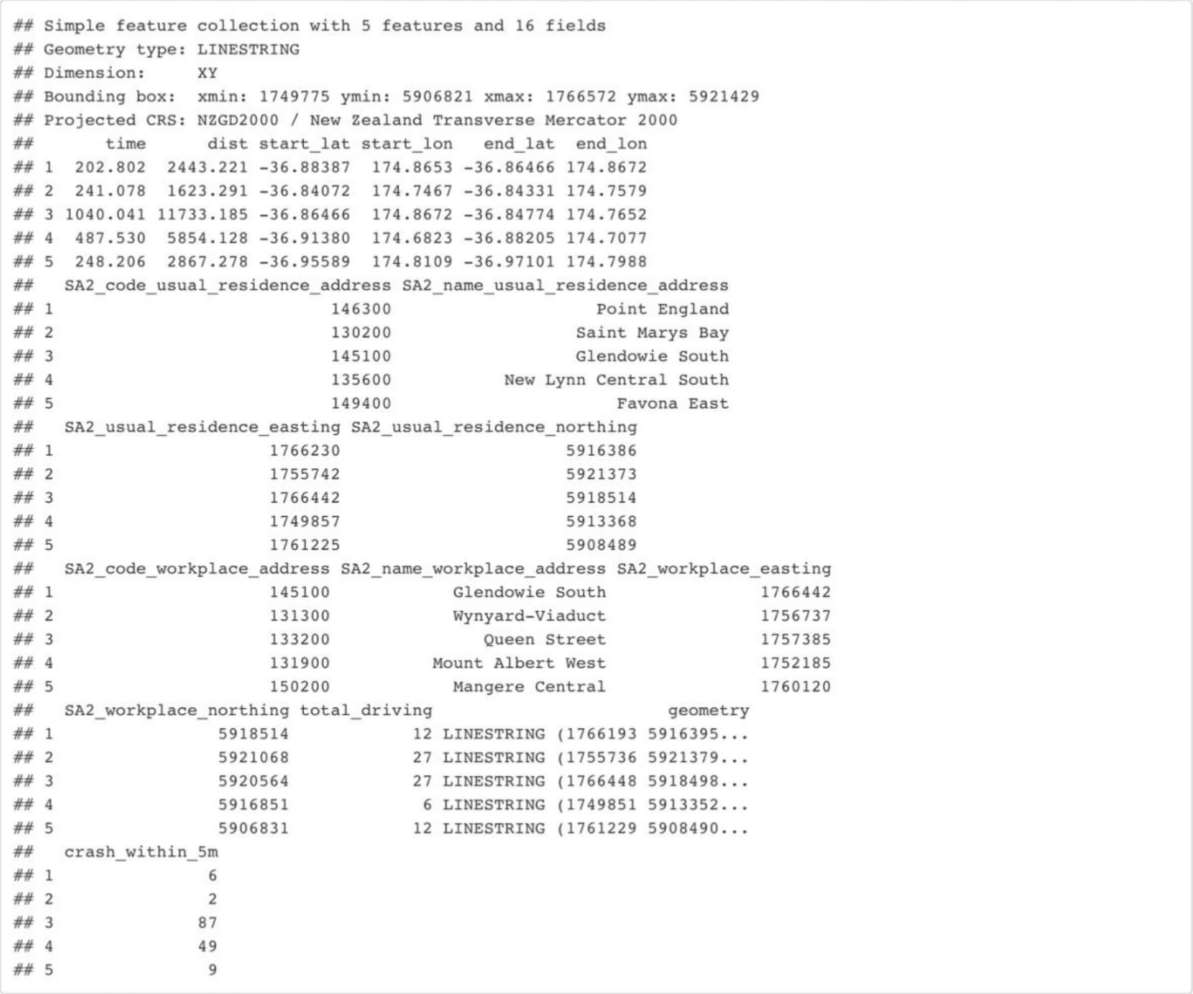


```
## Simple feature collection with 5 features and 10 fields
## Geometry type: GEOMETRY
## Dimension:     XY
## Bounding box:  xmin: 1710920 ymin: 5968143 xmax: 1763664 ymax: 6010086
## Projected CRS: NZGD2000 / New Zealand Transverse Mercator 2000
##   SA22021_V1          SA22021__1          SA22021__2 LAND_AREA_ AREA_SQ_KM
## 1     109700             Kaiwaka             Kaiwaka 259.022447 259.022447
## 2     109800     Mangawhai Rural     Mangawhai Rural  91.455113  91.455113
## 3     110200 Okahukura Peninsula Okahukura Peninsula 211.171451 211.171451
## 4     110400         Cape Rodney         Cape Rodney 370.681110 370.681110
## 5     110500           Wellsford           Wellsford   4.552902   4.552902
##   Shape_Leng     total_dist weight total_crash_weight              risk
## 1  204867.54   408188.0 [m]     42                 NA           NA [1/m]
## 2   89266.40   487640.1 [m]     48                 NA           NA [1/m]
## 3  202849.72 13384769.3 [m]   2364                  1 7.471178e-08 [1/m]
## 4  163226.49 22634880.7 [m]   2973                  8 3.534368e-07 [1/m]
## 5   11565.11  7654875.8 [m]   1545                  2 2.612714e-07 [1/m]
##                        geometry
## 1 MULTIPOLYGON (((1714038 598...
## 2 POLYGON ((1735474 5998474, ...
## 3 MULTIPOLYGON (((1714645 597...
## 4 MULTIPOLYGON (((1732289 598...
## 5 POLYGON ((1735673 5983601, ...
```

The *route_risk* function calculates the riskiness of a commuters' journey. The function requires two arguments the routes and crash data. Below is an example of how the *route_risk* function is used calculate the route risk using the *jtw_proj* as the routes and *cas_data_akl_2018_proj* as the crash data.

```
## getting route riskiness
rr_proj <- route_risk(jtw_proj, cas_data_akl_2018_proj)
```

The *route_risk* function returns the same information from the routes data frame with a new column called *crash_within_5m*, derived from counting the total number of crash sites a given route has encountered within five a meters radius. The radius default is 5 meters taking into consideration road width. The radius argument can be changed.

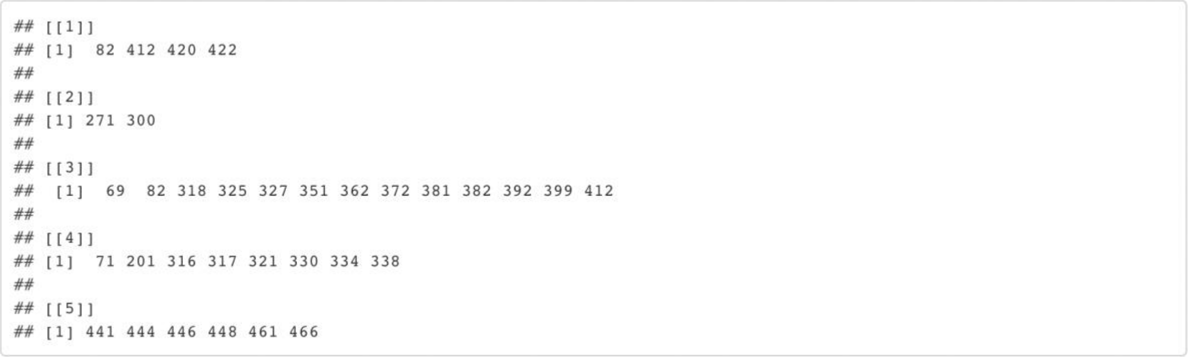


```
## Simple feature collection with 5 features and 16 fields
## Geometry type: LINESTRING
## Dimension:     XY
## Bounding box:  xmin: 1749775 ymin: 5906821 xmax: 1766572 ymax: 5921429
## Projected CRS: NZGD2000 / New Zealand Transverse Mercator 2000
##       time      dist start_lat start_lon   end_lat  end_lon
## 1  202.802  2443.221 -36.88387  174.8653 -36.86466 174.8672
## 2  241.078  1623.291 -36.84072  174.7467 -36.84331 174.7579
## 3 1040.041 11733.185 -36.86466  174.8672 -36.84774 174.7652
## 4  487.530  5854.128 -36.91380  174.6823 -36.88205 174.7077
## 5  248.206  2867.278 -36.95589  174.8109 -36.97101 174.7988
##   SA2_code_usual_residence_address SA2_name_usual_residence_address
## 1                           146300                    Point England
## 2                           130200                  Saint Marys Bay
## 3                           145100                  Glendowie South
## 4                           135600           New Lynn Central South
## 5                           149400                      Favona East
##   SA2_usual_residence_easting SA2_usual_residence_northing
## 1                     1766230                      5916386
## 2                     1755742                      5921373
## 3                     1766442                      5918514
## 4                     1749857                      5913368
## 5                     1761225                      5908489
##   SA2_code_workplace_address SA2_name_workplace_address SA2_workplace_easting
## 1                     145100            Glendowie South               1766442
## 2                     131300            Wynyard-Viaduct               1756737
## 3                     133200               Queen Street               1757385
## 4                     131900           Mount Albert West               1752185
## 5                     150200            Mangere Central               1760120
##   SA2_workplace_northing total_driving                       geometry
## 1                5918514            12 LINESTRING (1766193 5916395...
## 2                5921068            27 LINESTRING (1755736 5921379...
## 3                5920564            27 LINESTRING (1766448 5918498...
## 4                5916851             6 LINESTRING (1749851 5913352...
## 5                5906831            12 LINESTRING (1761229 5908490...
##   crash_within_5m
## 1               6
## 2               2
## 3              87
## 4              49
## 5               9
```

The *get_route_intersects* function identifies the polygons that have intersected with a given route. The function requires the argument 'routes' and 'polygons'. The routes argument is an *sf data frame* with an *sf linestring* geometry type and polygons is an *sf* data frame with a multi polygon geometry type. In the example below, *jtw_proj* and *sa2_clipped_proj_akl* are used as routes and polygons, respectively. The function also has an optional argument called inverse, where the default value is *False,* which refers to providing the polygon indices that the routes intersected as the function's output. Alternatively, if the 'inverse' argument is *True*, then the function returns the indices of the routes that intersected with the polygon.

```
## getting the intersects of each route
i <-  get_route_intersects(jtw_proj, sa2_clipped_proj_akl)
```

Below is a sample of the results of the *get_route_intersects* function. The function returns a list of the same length as the target variable, i.e., routes since the inverse is set to *False*. Each list element is a vector containing the indices of the SA2 polygons that intersected with *the jtw* routes.



```
## [[1]]
## [1]  82 412 420 422
##
## [[2]]
## [1] 271 300
##
## [[3]]
##  [1]  69  82 318 325 327 351 362 372 381 382 392 399 412
##
## [[4]]
## [1]  71 201 316 317 321 330 334 338
##
## [[5]]
## [1] 441 444 446 448 461 466
```

Chapter 7
Conclusion

7.1 Achievements

In this dissertation, we have presented that the developed R package has enabled us to conduct road safety analysis.

We estimated the routes for each journey using the 2018 census main means to work data which was subsequently used to calculate the distance travelled exposure and the road safety risk. The estimated routes were also used to quantify the riskiness of the routes. The traffic volume from estimated courses was compared with traffic volume from Auckland Transport's traffic counter. Moreover, spatial autocorrelation was used to conduct geospatial data analysis.

From the results of our analysis, we have concluded that there is a relationship between the distance travelled by the commuter and the likelihood of being exposed to a crash, such that the longer the distance travelled, the higher the probability that the commuter will encounter a traffic accident. Also, the presence of the 'close to home' effect on road accidents such that on shorter distance commute cannot be neglected on being exposed to traffic accidents. The comparison between the traffic volume from the simulated and traffic counter revealed a significant discrepancy because the routes were simulated using Statistical Area 2 centroids.

Moreover, from the spatial autocorrelation, we have identified several clusters south of Auckland and West Auckland where a commuter's journey from its residence is riskier than in other areas in Auckland. This can be attributed to commuters residing in South Auckland having a long journey to work. The algorithm prefers taking busy roads such as motorways, where many traffic accidents occur.

7.2 Limitations

The research focuses on road safety in the Auckland region of New Zealand. The limitation of this research is mainly from the datasets used. The 2018 census *main means of work* has been aggregated to the coordinates of the usual residence and workplace to centroids of statistical area 2, resulting in discrepancies on the simulated routes as it does not reflect the actual journey commuter has taken, thus excluding pertinent routes. Moreover, the crash dataset does not contain information on a specific day or period when the crash occurred, meaning we assume that the crashes occur during a weekday and peak hours when most journey to work occurs. Also, various factors that influence the journey have not been considered in this dissertation, such as traffic flow and speed variability.

7.1 Future Direction

The 2018 census main means to work data is aggregated into statistical geographic units that have reduced the quality of route simulation. It would be useful for a probabilistic approach to simulate routes in either statistical area 1 or mesh block level geographic units to improve the quality of the simulated routes. It would also be helpful to compare routes between the 2018 census with the upcoming *2023 census main means to work* data and identify the differences between the patterns and the proportions of the commuter by the mode of transport, given that the COVID pandemic has changed the dynamics of work environment [5]. It would also be worthwhile to understand to include speed and dissect the road into different parts, for example, road junctions, motorways, and suburban roads, to understand its relationship with traffic accidents throughout New Zealand.

References

[1] New Zealand Government (2019). *New Zealand's Road Safety Strategy 2020-2030.* https://www.transport.govt.nz/assets/Uploads/Report/Road-to-Zero-strategy_final.pdf

[2] World Health Organization. (2018). Global status report on road safety 2018: Summary (No. WHO/NMH/NVI/18.20). *World Health Organization*.

[3] Wegman, F. (2017). The future of road safety: A worldwide perspective. *IATSS research*, *40*(2), 66-71.

[4] Wickham, H. (2015*). R packages: organize, test, document, and share your code.* " O'Reilly Media, Inc.".

[5] Institute for Transport Studies (ITS) University of Leeds. (n.d.). *trafficalmr.* Retrieved 1July 2021 from https://saferactive.github.io/trafficalmr/

[6] Nævestad, T. O., Phillips, R. O., & Elvebakk, B. (2015). Traffic accidents triggered by drivers at work–A survey and analysis of contributing factors. *Transportation research part F: traffic psychology and behaviour*, *34*, 94-107.

[7] Scuffham, P. A., & Langley, J. D. (2002). A model of traffic crashes in New Zealand. *Accident Analysis & Prevention*, *34*(5), 673-687.

[8] Belin, M. Å., Tillgren, P., & Vedung, E. (2012). Vision Zero–a road safety policy innovation. *International journal of injury control and safety promotion*, *19*(2), 171-179.

[9] Wegman, F., & Oppe, S. (2010). Benchmarking road safety performances of countries. *Safety science*, *48*(9), 1203-1211.

[10] Elvik, R. (2009). The trade-off between efficiency and equity in road safety policy. *Safety science*, *47*(6), 817-825.

[11] Wong, S. C., & Sze, N. N. (2010). Is the effect of quantified road safety targets sustainable?. *Safety Science*, *48*(9), 1182-1188.

[12] Hakkert, A. S., Braimaister, L., & Van Schagen, I. (2002). *The uses of exposure and risk in road safety studies* (Vol. 2002, No. 12). SWOV, Leidschendam: SWOV Institute for Road Safety.

[13] Leur, P. D., & Sayed, T. (2002). Development of a road safety risk index. *Transportation Research Record*, *1784*(1), 33-42.

[14] Shah, S. A. R., Ahmad, N., Shen, Y., Pirdavani, A., Basheer, M. A., & Brijs, T. (2018). Road safety risk assessment: an analysis of transport policy and management for low-, middle-, and high-income Asian countries. *Sustainability*, *10*(2), 389.

[15] Pei, X., Wong, S. C., & Sze, N. N. (2012). The roles of exposure and speed in road safety analysis. *Accident analysis & prevention*, *48*, 464-471.

[16] Merlin, L. A., Guerra, E., & Dumbaugh, E. (2020). Crash risk, crash exposure, and the built environment: A conceptual review. *Accident Analysis & Prevention*, *134*, 105244.

[17] Burdett, B. R., Starkey, N. J., & Charlton, S. G. (2017). The close to home effect in road crashes. *Safety science*, *98*, 1-8.

[18] Lovelace, R., Nowosad, J., & Muenchow, J. (2019). *Geocomputation with R.* Chapman and Hall/CRC.

[19] Stal, C., De Sloover, L., Verbeurgt, J., & De Wulf, A. (2022). On Finding a Projected Coordinate Reference System. *Geographies*, *2*(2), 245-257.

[20] Getis, A. (2010). Spatial autocorrelation. In *Handbook of applied spatial analysis* (pp. 255-278). Springer, Berlin, Heidelberg.

[21] Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological modelling*, *135*(2-3), 147-186.

[22] Manap, N., Borhan, M. N., Yazid, M. R. M., Hambali, M. K. A., & Rohan, A. (2021). Identification of hotspot segments with a risk of heavy-vehicle accidents based on spatial analysis at controlled-access highway. *Sustainability*, *13*(3), 1487.

[23] Bivand, R. (2022). R Packages for Analyzing Spatial Data: A Comparative Case Study with Areal Data. *Geographical Analysis*.

[24] Waka Kotahi New Zealand Transport Agency. (n.d.). Retrieved 1 July 2021 from https://www.nzta.govt.nz/safety/partners/crash-analysis-system/

[25] Ministry of Transport. (n.d.). Retrieved 1July 2021 from https://www.transport.govt.nz/area-of-interest/safety/road-to-zero/

[26] Blick, G., & Donnelly, N. (2016). *From static to dynamic datums: 150 years of geodetic datums in New Zealand.* New Zealand Journal of Geology and Geophysics, 59(1), 15-21.

[27] Statistic New Zealand. (2020, June 14). https://datafinder.stats.govt.nz/table/104720-2018-census-main-means-of-travel-to-work-by-statistical-area-2/

[28] Auckland Transport. (n.d.). Retrieved 1 July 2021 from https://at.govt.nz/about-us/reports-publications/2018-census/.

[29] Statistic New Zealand. (2021, July 05). https://catalogue.data.govt.nz/dataset/statistical-area-2-2018-clipped-generalised

[30] Auckland Transport. (March 2022). https://at.govt.nz/about-us/reports-publications/traffic-counts/